Improving Peer Feedback in Project-Based Learning Contexts: An Investigation into a First-Year Engineering Intervention

Ms. Katherine Drinkwater, Virginia Polytechnic Institute and State University

Katie Drinkwater is a first-year PhD student in Engineering Education at Virginia Tech. She received a Bachelor's degree in Mechanical Engineering from Duke University. Her research interests include engineering extracurriculars, PBL, design in informal learning environments, makerspaces, and women in engineering.

Olivia Ryan, Virginia Polytechnic Institute and State University

Olivia Ryan is a Ph.D. student in Engineering Education at Virginia Tech. She holds a B.S. in engineering with a specialization in electrical engineering from Roger Williams University. Her research interests include developing professional skills for engineering students and understanding mathematics barriers that exist within engineering.

Marin Jayne Fisher Hale, Virginia Polytechnic Institute and State University

Marin is a doctoral student from Aurora, IL. She holds a B.S. in mechanical engineering from Brigham Young University. Her current research interests include teamwork in student teams and religious identity within engineering. When not working on research, she enjoys hikes, road biking, and coaching high school color guard.

Susan Sajadi, Virginia Polytechnic Institute and State University

Susan Sajadi is an assistant professor at Virginia Tech in the department of engineering education. She has a BS and MS in Biomedical Engineering and a Ph.D. in Engineering Education Systems and Design from Arizona State University. Prior, she worked as an engineer in the medical device industry.

Dr. Mark Vincent Huerta, Virginia Polytechnic Institute and State University

Mark Huerta is an Assistant Professor in the Department of Engineering Education at Virginia Tech. He earned his PhD in Engineering Education Systems & Design at Arizona State University and has a BS/MS in Biomedical Engineering. His research focuses on exploring and understanding engineering learning environments. He harnesses these insights to propose solutions that encourage the creation of safe and inclusive educational environments conducive to learning, professional development, and innovation. His research interests include graduate student mentorship, faculty development, mental health and well-being, teamwork and group dynamics, and the design of project-based learning classes.

Improving Peer Feedback in Project-Based Learning Contexts: An Investigation into a First-Year Engineering Intervention

Abstract

In an increasingly collaborative and globalized world, effective teamwork is an essential skill for engineers. Peer evaluation is a valuable practice that helps students assess and improve their teamwork skills, especially in project-based learning (PBL) courses. While popular peer evaluation tools like CATME collect both quantitative ranking and qualitative comments, qualitative peer comments often lack objective, helpful feedback due to several potential biases. In this paper, we describe the implementation and impact of a 30-minute interactive presentation intended to teach first-year engineering students to give and receive quality feedback.

To further investigate the effectiveness of the in-class intervention on peer feedback quality, a rubric was adapted to measure the quality of peer feedback comments. Preliminary findings show significant improvement in all criteria for feedback quality (Task, Gap, and Action) for students who received the intervention, with the largest gain in students writing peer comments with more actionable feedback We also found a significant difference in the length of peer feedback comments between the class with the intervention and the class without the intervention. However, throughout data analysis, we observed gaps in our chosen framework, and as such, we are developing and testing an improved rubric to quantitatively rate student feedback. This paper will help instructors learn an approach toward aiding students in writing actionable feedback, improving the overall quality of qualitative peer comments. Further, we present the development of a rubric that can be used to assess peer feedback, which can further the understanding and impact of formative peer feedback in first-year engineering courses.

I. Introduction

In an increasingly collaborative and globalized world, effective teamwork is an essential skill for engineers [1]. To help students develop teamwork skills, project-based learning (PBL) courses, including first-year cornerstones, have become a component of most engineering programs [2]. However, having students work in teams on an engineering project does not necessarily guarantee effective teamwork is practiced or that students further develop their teamwork skills [3]. Peer evaluation systems, such as Comprehensive Assessment of Team Member Effectiveness (CATME), have been developed to help instructors monitor team dynamics and better support teams by assessing teamwork skills and providing formative feedback to team members throughout the course [4]. Formative feedback provided via peer evaluation has been shown to encourage effective team behaviors and decrease social loafing [5], [6].

While CATME evaluations collect both quantitative rankings and qualitative comments, qualitative peer comments often lack objective, helpful feedback due to several potential biases [7]. These biases can arise from personal connections between teammates, a team member's gender, or fear of harming the team dynamic. To combat these challenges, it has become clear

that intentional instruction on how to give feedback is essential [8], [9], [10]. Previous literature shows an improvement in the quality of quantitative CATME ratings after students received frame-of-reference training [11], [12]. Huang et al. [10] found an increase in the number of CATME dimensions addressed in qualitative comments after conducting an intervention with senior capstone students that more clearly defined each CATME dimension. Informed by the positive results of previous interventions and our own challenges with low-quality written feedback, we explored the implementation of a feedback intervention for engineering students in a first-year PBL course. We created a 30-minute, interactive presentation that covered common pitfalls of peer feedback, qualities of constructive feedback, and examples of helpful and ineffective feedback. Our intervention was presented to six classes of first-year engineering students at a large, public university in the Mid-Atlantic region of the U.S..

This paper highlights the initial steps of a larger study that seeks to understand the impact of a feedback intervention on peer feedback quality. The potential impact can be elucidated by comparing comments from students who received the intervention with students from a previous year who did not. This paper describes the development of a rubric to assess peer feedback quality that arose from the implementation and evaluation of the intervention. To examine this process, we ask the following research questions:

1. How did exposure to a feedback intervention in a first-year engineering course impact the quality of peer feedback comments?
2. What are the components of quality peer feedback, and how effective are existing rubrics in measuring the quality of peer feedback comments?

II. Background

A. Role of Feedback in PBL Courses

Project-based learning (PBL) courses are a common pedagogical approach used to teach engineering design [13], especially in senior capstone and first-year cornerstones. The team- and project-focused nature of PBL courses helps students develop essential professional skills such as communication [14], conflict management [15], and collaboration with diverse team members [16]. Another unique aspect of the PBL format is the team dynamics in every project group. Each team forms a culture and workflow unique to their team, which can help or hurt the team's productivity. The course instructor is not involved in most team interactions and, thus, is less equipped to judge the influence of individual students on team dynamics. Peer evaluation tools fill this gap by eliciting feedback from the people most familiar with the team (i.e., team members). This process informs the instructor about team dynamics and helps teams improve their dynamics and performance [17].

To utilize peer evaluation opportunities to improve team performance and reflect on areas of individual growth, students must be familiar with desirable teamwork behaviors and must be able to clearly communicate constructive feedback to their peers. Unfortunately, it is rare for peer

feedback to address teamwork behaviors in a detailed and actionable manner [18]. With minimal guidance from instructors, peer feedback more often tends toward short phrases like “good teammate” or personality-based comments that do not relate to the team (e.g., “super funny person”, “great at fantasy football”). This warrants the need to improve feedback quality, which can improve team performance and student skill development.

B. Existing Peer Feedback Interventions in Engineering Education

Several researchers have also acknowledged the lack of feedback quality in peer evaluations–both quantitative and qualitative. Prior interventions in engineering education contexts teach students how to give feedback and rate their peers. Loignon et al. [11] trained student groups in frame-of-reference (FOR) before the groups completed a bridge-building activity. FOR training “emphasized the multidimensionality of performance” by defining each CATME dimension and providing examples of high and low scores in each dimension. The students rated their teammates in CATME after the activity, and students who received the training intervention showed more consistent rating (decreased rater variance) and greater variance between teammates (increased target variance) than students who did not receive the training. Building upon Loignon and colleagues’ work, Mertz, Ferguson, and Hoque [12] implemented FOR training and rater error training in an Intro to Engineering course. The rater error training informed students of common rating problems (e.g., bimodal rating, giving everyone the same score) and suggested solutions to differentiate between team members’ performance. Classes that received both interventions had better alignment between self and peer CATME ratings and greater variance of scores across the class. Trainings like FOR and rater error help with quantitative peer ratings but do not provide much guidance for qualitative peer comments. Recently, Huang and colleagues [10] adapted the five CATME quantitative dimensions (Contributing to the Team's Work, Interacting with Teammates, Keeping the Team on Track, Expecting Quality, and Having Relevant Knowledge, Skills, and Abilities) into a training intervention for qualitative comments. The researchers taught students about behaviors that represent each dimension and how to evaluate a team member’s performance holistically. Results showed a significant increase in the number of CATME dimensions mentioned in qualitative peer comments.

These feedback interventions are encouraging examples that show how teaching students to appropriately rate and describe peer performance can improve student feedback results. However, we could not identify an existing example of an intervention focused on actionable feedback comments or an intervention independent of the evaluation tool (i.e., CATME) within engineering education contexts. Thus, we sought to fill this gap with our intervention.

C. Theoretical Frameworks and Existing Rubrics

To evaluate the impact of the feedback intervention on students' peer comments, we researched frameworks and rubrics to understand what constructive, effective peer feedback looks like and

how peer feedback is normally evaluated. We found that feedback rubrics and theoretical frameworks fall along a spectrum of complex to general, meaning some rubrics are looking for very specific requirements, while others consider broad themes to evaluate feedback.

Hattie and Timperly present a well-known feedback framework for effective feedback. Their framework includes four levels of feedback: task level, process level, self-regulation level, and self level [19]. We considered using this framework to evaluate the student comments; however, we found it difficult to differentiate between the levels for relatively brief peer comments, so we considered other models. In our search of frameworks and rubrics, we observed that a significant amount of the peer feedback literature is based in medical education because common pedagogical practices in the field, such as clinical rotations, necessitate immediate feedback from an instructor. Still, their approaches can likely be applied to other group or team-based learning settings. A well-known approach to evaluating teamwork behaviors includes the TeamUP rubric. The TeamUP rubric was developed in the form of a Likert-scale survey for midwifery students and includes five domains: planning, environment, facilitating the contributions of others, managing conflict, and contributing to the team project [20]. Although we were not planning to have students fill out a survey for peer assessment, the domains in which they considered peer assessment were valuable in our understanding of peer feedback. Another medical education paper shared a model of peer feedback for clinical skills assessments. The model included six domains: control of syntax and mechanics, quality of comments, balance of comments, positive feedback phrasing, negative feedback phrasing, and appropriate suggestions [21]. Although this model is thorough, we felt that it was too focused on the delivery style for our context.

The complexity of some of the other feedback rubrics felt challenging to implement on relatively short feedback comments, so we considered rubrics that approached evaluating feedback more generally. Gauthier et al. from the medical education field developed a rubric using the Task, Gap, Action framework originally presented by Sadler. While Sadler's work was concerned with formative assessment in all educational fields [22], the rubric created by Gautheir et al. evaluates peer feedback for residency training across three criteria. The task describes the context in which the feedback was given, which can include the content of a student's participation or the value that their participation added to the team. The gap recognizes the differences between behaviors displayed in a comparative measure, and the action describes what can be improved [23]. Other studies in medical education have used the Gauthier et al. rubric to evaluate undergraduate students [24] and medical students in a team environment [18].

We chose the Task, Gap, Action framework to evaluate peer comments for our study because of its simplicity and adaptability to our context. The peer comments we collected were relatively brief, so it was impractical to evaluate the comments on more than five criteria; therefore, we chose to use the simpler framework for our project-based learning engineering context. In general, one of the challenges we faced when selecting a rubric/framework was the applicability

to our context. The majority of the rubrics and framework were developed for instructor-to-student feedback, but we evaluated student-to-student feedback. This application likely has similarities, but the content and presentation of feedback an instructor provides will be different from the feedback a student provides. Therefore, we sought to test the fit of this existing rubric in the first-year engineering context.

III. Methods

A. Sample

This study is part of a larger project to examine peer feedback in an engineering PBL context. The subset of data presented here includes four course sections from one instructor over two years. Two sections from Fall 2022 were included as a control group, and two of the six sections that participated in the intervention in Fall 2023 were included. The demographics of the 87 participants from Fall 2022 and 118 participants from Fall 2023 are shown below in Tables 1 and 2. Each section contained approximately 70 students. The instructor, class times, and duration were the same in both years. The distribution of genders and races/ethnicities is fairly consistent between years. More than 70% of the participants were Male in both years, and White and Asian students make up more than half of the races/ethnicities selected, which is similar to the engineering population at our university.

TABLE I
Gender of Study Participants by Year

	Male	Female	Other or Prefer not to answer	Total
Fall 2022	64	17	6	87
Fall 2023	86	26	6	118

TABLE II
Race or Ethnicity of Study Participants by Year

Race or Ethnicity	Number of Students F22	Number of Students F23
Asian	28	40
Black or African American	4	8
Hispanic, Latino, or Spanish Origin	9	14
Middle Eastern or North African	0	2

Native Hawaiian or other Pacific Islander	1	0
White	29	40
Other or Prefer not to Answer	16	14
Total	87	118

B. Course and Intervention

This study was conducted within the second course of a first-year engineering course sequence running in the off-cycle. Every engineering student must complete the two introductory courses before declaring an engineering major. The focus of the course is a team-based design project in which students navigate an engineering design process and apply 3D modeling, Arduino programming, and prototyping tools to create an integrated prototype.

Activities and assessments were implemented throughout the course to support collaboration and mitigate social loafing. A key component of this was the use of peer evaluation. In both the Fall 2022 and 2023 semesters, the CATME peer evaluation system was used to elicit and facilitate feedback, which included collecting student quantitative responses to key teamwork behaviors and qualitative self and peer comments.

1. Intervention in Fall 2023

To improve the quality of comments students provided in CATME, an intervention was designed and implemented in the Fall 2023 semester focused on providing students with guidance on how to write effective feedback. The lecture used an interactive presentation software, Mentimeter, to foster student engagement and promote discussions. The primary emphases of the training were to 1) introduce students to CATME, including the quantitative rating scales and comment boxes, 2) discuss student concerns about the peer evaluation process, and 3) provide students with tips on how to write and receive peer feedback comments based on good feedback practices from the literature.

After providing an overview of CATME, the instructor posed several open-ended questions asking students why they think instructors use peer evaluations and whether they have concerns about the process. The purpose of fostering this discussion was to help students understand the purpose of peer evaluation: provide one another feedback and help the instructor monitor team dynamics and identify potential issues. Additionally, the instructor addressed apprehensions about peer evaluations, such as not wanting to negatively impact someone else's grade or causing contention in a team.

Informed by the Task, Gap, Action framework, the instructor then provided guidance to students on how to write effective feedback. The importance of specific feedback grounded in observable teamwork behaviors (Task) was highlighted. A specific emphasis was also placed on the value of providing both positive, reaffirming feedback as well as constructive, actionable feedback (Gap and Action). Specific examples of feedback from previous semesters were used to help students understand and better discern high-quality feedback versus poor, generic statements. For example, a helpful comment is "XXXX made sure everyone stayed focused on the project. An example is when he initiates texts in the groupchat on when things are due and what he is working on.", but a poor example is "XXXX gets his work done". Students were instructed on how to provide constructive criticism and asked to alter a comment to make it more actionable. Finally, the instructor gave tips on receiving feedback. He described how it is normal to feel defensive or angry, but humbly accepting feedback and taking accountability is essential for professional growth.

We observed that students were engaged in the interactive presentation and interested in writing and receiving better feedback. The intervention addressed every step of the process (quantitative rating, qualitative comment writing, receiving feedback, reflecting, and taking action) in an effort to make first-year students more comfortable with and cognizant of the peer evaluation process.

C. Data Collection

Students were tasked with providing peer feedback on CATME through quantitative ratings and qualitative comments on teamwork behaviors three times during the semester. They were given instructions to comment on at least two specific elements on team behavior for each team member (including themselves) and were encouraged to provide both positive and constructive, actionable feedback. For both years of data included in this project, students gave feedback at three time points in a 16-week semester (Week 7, Week 11, and Week 16). The first and second CATME evaluations were used to analyze the difference between students in Fall 2022 (F22) and Fall 2023 (F23), the year the intervention was implemented. The third CATME evaluation focuses on summative feedback, so it was not included in our analysis as our focus was enhancing the quality of formative feedback. All CATME evaluation data was exported by the instructor and deidentified prior to analysis. This study was approved by the IRB at the authors' institution.

D. Data Analysis

To analyze the difference in feedback quality between first-year engineering students in 2022 who did not participate in the feedback intervention and 2023 students who did, we analyzed several metrics of comment quality. The first was an analysis of the length of comments, followed by a more in-depth coding of the comments using the Task, Gap, Action rubric adapted from Gauthier et al. [23].

1. Comment Length

The length of peer comments is a quantitative measure that allowed us to gain an initial understanding of the differences in peer comments between the two groups. We measured the number of words in each peer comment to determine length. While a longer comment length does not guarantee higher quality feedback, longer comments are typically necessary to describe a peer's performance in enough detail to be actionable. Through our initial review of the data, the research team noticed that the highest-quality comments tended to be the longest in length. Thus, we selected length as a quantitative measure to compare the two groups of students. Descriptive statistics and a Mann-Whitney U-Test were performed on the data [25].

2. Feedback Quality

For a deeper analysis of feedback quality, we graded each peer comment using a rubric adapted from Gauthier et al. [23] (Table 3). This rubric was selected because of its focus on opportunities for improvement (gaps) and actions to improve performance (action).

TABLE III
Initial Rubric of Feedback Quality, adapted from (Gauthier et al., 2015)

Feedback Element	0	1	2	3
Task (description of behaviors or tasks to the ratee upon which feedback was given)	No tasks described	Vague description; lacking either content of ratee's work or value contributed to the group	General description of content of task; value to team not included	Specific description of both content and value of the ratee's work
Gap (recognition of a difference--positive or negative--between the ratee's work and a standard)	No gap described	Gap(s) alluded to but not described (e.g. they did a great job of leading, missed a few classes).	Gap(s) generally described	Gap(s) and comparative standard specifically described
Action (Suggested steps to remedy gaps or improve performance)	No actions described	Action is recommended but no suggestions are described	Actions recommended in general terms but lack specificity to the team or project	Specific actions that are relevant to the team and project are recommended

The researchers performed one round of coding using this rubric. Two raters rated each peer comment 0-3 in each element (task, gap, action). Inter-rater reliability was calculated using a weighted Cohen's kappa [26]. Descriptive statistics and non-parametric Mann-Whitney U-Tests

[25] were calculated with the quantitative scores to assess whether there was a difference in the quality of feedback between the two years.

IV. Results

A. Comment Length

The comment length was evaluated to determine if there was a difference in how much students wrote about their peers with and without participating in a feedback intervention. Since the data was not normally distributed, we used the non-parametric Mann-Whitney U-Test to compare the number of words in students' peer comments from Fall 2022 and Fall 2023 classes. It was found that there was a statistically significant difference ($U = 47406$, $p < .001$) between the number of words used in peer comments between Fall 2022 and 2023. Additionally, the descriptive statistics in Table 4 show that the average number of words in student peer comments in Fall 2022 and Fall 2023 were 28.36 and 45.11 words, respectively. This suggests that, on average, students wrote longer peer comments after participating in the feedback intervention workshop. Furthermore, Figure 1 shows boxplots of the distribution of comment lengths between the two years. For both years, any outliers were comments that were much longer than the average comment. As can be seen in Figure 1, Fall 2023 has more longer outlier comments than Fall 2022. Again, this suggests that students wrote longer peer comments after participating in the feedback intervention workshop.

TABLE IV
Comment Length Descriptive Statistics

Year	n	Mean	Median	IQR	Standard Dev.	Skewness	Kurtosis
F22	372	28.36	25	25	19.71	1.24	4.92
F23	399	45.11	38	28	31.34	1.64	6.20

Note: n is the number of peer and self comments analyzed.

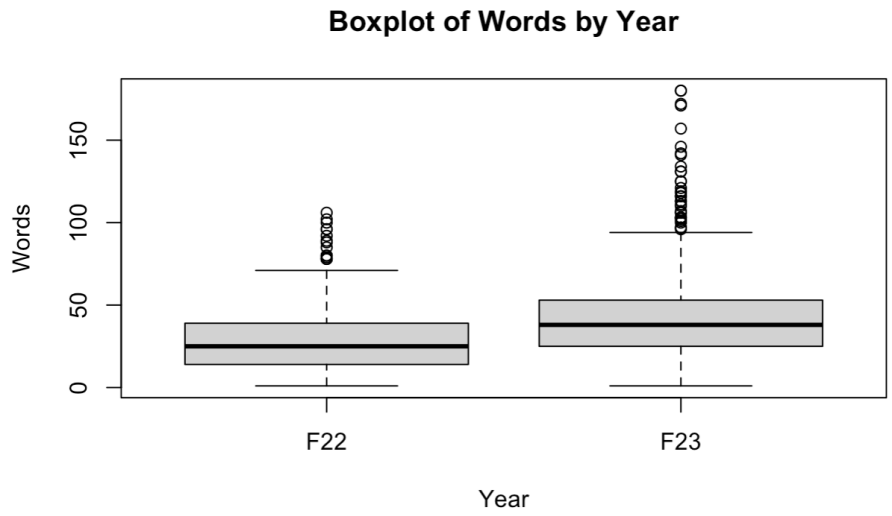

Fig. 1 Boxplots of Number of Words in Comment by Year

B. Feedback Quality

The comment quality data shows promising results. Each category from the rubric (task, gap, and action) was statistically significant in a Mann-Whitney U-Test [25]. The results were as follows: Task ($U = 58734$, $p <. 001$), Gap ($U = 54320$, $p < .001$), and Action ($U = 55367$, $p < .001$). The effect sizes (Task = .206, Gap = .264, Action = .25) are small, however, likely due to the wide range of comment quality and results of the inter-rater reliability calculations (discussed below in Table 5 and Table 6).

TABLE V
Comment Quality Rating Descriptive Statistics

Domain	Year	n	Mean	Median	Standard Dev.
Task	F22	372	1.184	1	0.92
	F23	399	1.471	1	0.889
Gap	F22	372	0.63	0	0.907
	F23	399	0.957	1	1.04
Action	F22	372	0.118	0	0.505
	F23	399	0.463	0	0.818

Note: n is the number of peer and self comments analyzed.

TABLE VI
Results for Comment Quality Rating

Independent Samples U-Test (Mann-Whitney)	Statistic	P-Value	Effect Size (Rank biserial correlation)
Task	58734	<.001	0.206
Gap	54320	<.001	0.264
Action	55367	<.001	0.25

1. Inter-Rater Reliability

After two researchers scored each peer comment based on the rubric, inter-rater reliability was calculated using a weighted Cohen's Kappa measure of agreement [26], shown in Table 7. Based on these findings, all codes had reliability scores well below the recommended value of 0.7, indicating limited agreement. The quadratically weighted kappa values are slightly higher than the linearly weighted values, which indicates that smaller disagreements in rating (one point difference) were more common than large disagreements (two or three point difference). Although the quadratically weighted kappa values were slightly better, these results highlighted ambiguities in the rubric we were using, which led us to iterate the Task, Gap, Action rubric we were using.

TABLE VII
Inter-rater Reliability Scores

Criterion	Linear Weighted Cohen's K	Quadratic Weighted Cohen's K
Task	0.37	0.52
Gap	0.05	0.05
Action	0.59	0.69

V. Discussion

Addressing Research Question 1, analysis of peer feedback comments performed in this study suggests that the intervention had a positive effect on feedback quality. There are encouraging trends within the data that show the benefit of the feedback intervention. This process, however, also revealed several gaps in current frameworks used to assess feedback, which allowed us to add Research Question 2 to our study. Overall, the challenges we experienced during data analysis and some reliability concerns limit the conclusions we can make from this data. We discuss our preliminary findings for the length analysis and feedback quality rating briefly below. Due to the addition of RQ2 and our work to iterate upon our original rubric, the remainder of this section focuses on limitations to the original rubric and our ongoing revisions. This research contributes to the current corpus of research on improving feedback quality in peer evaluation.

We join Huang et al. [10] in highlighting the value of qualitative comments to build strong teamwork skills in engineering students.

The comment length analysis showed a significant increase in the number of words used in Fall 2023 comments compared to the Fall 2022 students who did not receive the intervention. The mean comment length of F23 students is 59% higher than F22 students. The variance of F23 data is much larger ($\sigma = 31.34$ for F23, $\sigma = 19.71$ for F22) due to many outliers with very long comments. A larger distribution of comment length could suggest that students varied the length of comments depending on the student they were rating or could indicate that the intervention impacted some students more than others. This positive trend in the F23 group is also present in the feedback quality ratings, as there was a significant increase ($p < 0.001$) in the scores for all of the Task, Gap, and Action criteria in the F23 group. The mean scores of all three criteria increased by approximately 0.3 points, with Action showing the largest increase at 0.345 points. This result aligns with the feedback intervention, which focused on including actionable, constructive comments in peer evaluation. Yet, despite this increase, the mean scores for both years were still very low for all three criteria. No mean score exceeded 1.5, and the means for Gap and Action were below 1.0. This indicates that the majority of comments did not address a performance gap or action at all. Aspects of the rubric design also had an influence on these relatively low scores. Thus, there is still room for significant improvement in peer feedback quality and refinement of feedback assessment tools.

A. Limitations

The rubric adapted from Gautheir et al. (Table 3) was selected for its simplicity, focus on action, and ability to quantitatively grade feedback comments. However, as we progressed through data analysis, the research team realized that using this rubric was not as straightforward as we had hoped. The poor inter-rater reliability scores (Table 7) forced us to evaluate inconsistencies with the application of the rubric. We identified problems with defining the Task and Gap criteria and distinguishing between scores, which led to limited use of the current feedback rubric. In the following sections, we highlight the challenges with the Task, Gap, Action framework as well as limitations of the study context.

1. Defining Task

For the Task criterion, we realized that peer comments described two main types of task performance: performance on tasks specific to this course's project and performance related to general self-regulation and teamwork behaviors. A peer could describe another student's integral job of coding the Arduino before the prototype was due (specific task), or they could describe how a student's organizational skills were essential for the team (behavior). Based on the definition of the Task criterion, there was no place in the rubric to grade comments that addressed behaviors. An additional problem with the Task criterion was the inclusion of both content and value of the task. To score a '3' in the Task criterion, a comment needed to describe

both the content of the task and the value added to the team. This was problematic in cases where the comment described the content in detail, but the value was implied. It was often easy for the raters to see how many positive tasks and behaviors add value to the team, even without explicit mention of the value.

2. Defining Gap

For the Gap criterion, raters differed on whether performance above the comparative standard counted as a gap. When initially defining the rubric, two authors envisioned the Gap domain as describing feedback where a student was either performing below the team standard (negative gap) or above the standard (positive gap), while three authors solely considered a gap as negative. Additionally, when coding the data, we ran into many instances where the performance standard was unclear. If a student commented on their peer with the following: "misses class occasionally and late to team meetings, but does have drive to get work done and on time," are they implying that the peer is not meeting the standard of attendance? For those that included both positive and negative gaps, it was difficult to consistently grade comments because the relationship between Gap and Action was muddled. It is easy to see how a negative Gap and corrective Action are connected, but what is the action that accompanies a positive gap? Do students need to explicitly tell their peers to keep performing well? The blurry connection between a performance standard, the gap, and the requisite action made it difficult to rate comments on the Gap criterion.

3. Distinguishing Between Numerical Scores

Our points of confusion with the Action criterion centered around the lack of variation in scores. We mostly gave zeros, and almost never graded comments with a three, even for comments we would anecdotally consider good feedback. This problem was most pronounced in the Action criterion, but the scores were low for all three criteria. This could be attributed to low-quality feedback in the peer comments or inconsistencies in rating. We often had difficulty distinguishing between a one or two score. In the rubric (Table 3), comments with a vague description should be scored with a one, while comments with a general description should score a two. Distinguishing between general and vague descriptions proved very difficult. The lack of score distribution forced us to reconsider whether the rubric was appropriate for this data. Gauthier et al. developed their rubric to evaluate feedback from instructors to medical students. An instructor is more likely to know and suggest corrective action, especially in the case of a patient's treatment in a clinical rotation. Engineering students in a PBL context may envision an action they wish their teammates would make, but a peer lacks the authority that an instructor possesses to dictate the action. The unique relationship behind peer feedback necessitates a different kind of feedback. Based on the challenges, inconsistencies, and questions we identified during data analysis, we conclude that the rubric used to grade peer feedback quality must be iterated upon to accurately capture the important elements of peer feedback in an engineering

education context. In the next section, we expand upon our ongoing development of a new rubric.

4. Limitations due to Study Context

Challenges and inconsistencies with the rubric may have compounded other limitations of this study. The intervention highlighted in this study was developed as part of a larger project that uses peer feedback comments and generative AI to create Performance Feedback Reports (PFRs) for students [27]. Students in the Fall 2023 classes received a PFR between the first and second CATME evaluations. The experience of receiving a comprehensive feedback report could have impacted how F23 students wrote their peer comments for the second CATME evaluation. While many of the contextual factors were constant between years, the samples from both years exhibited a high percentage of male students. This is not uncommon for an engineering context, but it still limited the representation of comments written by or for non-male students within the sample.

B. Future work

After the initial application of the rubric and identification of the many limitations to our current approach, we recognized that we needed to improve and clarify the rubric to better reflect the type of comments in our data. The authors collaborated to further clarify the criteria and score levels of the rubric and add examples for each score. In this section, we present initial changes to the rubric and describe our ongoing future work to validate the refined rubric.

In the team-based PBL environment of this study, we believe that behaviors and tasks are two separate but important elements of teamwork and cannot be rated within the same criterion. Therefore, we separated the Task criterion into Contribution to Group Tasks and Behavior. We renamed the Task criterion to better capture all of the ways that students contribute to the team. A student can work diligently on tasks that do not directly contribute to an upcoming assignment but still add value to the team. Considering contribution to tasks also inherently incorporates the value of the task. The description of each score in the Contribution to Group Tasks and Behavior criteria is now based on a description of the content *or* value of the student's actions. This removes the confusion caused by needing content *and* value to score highly in the Task criterion. Activities included in the Contribution to Group Tasks criterion include role fulfillment, task management, task execution, project contributions, and work output. The Behavior criterion considers actions that are not tied to the current project, like interpersonal dynamics, team engagement, communications, or personal attributes. Differentiating between tasks and behaviors will help us clarify our ratings by defining a specific criterion to rate behavior-focused comments. Additionally, this differentiation aligns with the task and self-regulation levels of feedback in Hattie and Timperly's model of feedback [19]. Hattie and Timperley assert that providing feedback on self-regulation behaviors and task performance is key to long-term performance improvement.

To remedy confusion between positive and negative gaps in the Gap criterion, we limited the Gap criteria to only focus on negative gaps. Positive gaps are included in the ratings for the Task and Behavior criteria. This restores the connection between the Gap and Action criteria and allows the researchers to focus on the presence of constructive feedback in peer comments. With this change, we realized that the Task and Behavior criteria focus primarily on reinforcing positive performance, while the Gap and Action criteria focus on constructive feedback to improve performance. This finding aligns with what we taught in the feedback intervention; both affirming and constructive comments are important for quality feedback.

Because the scoring problems were present in every criterion, we decided to collapse the '1' and '2' scores into the same score. By changing the scale from 0-3 to 0-2, we assert that there is not a significant increase in the quality of a 'general' comment compared to a 'vague' comment. This change allowed us to delineate the difference between scores more clearly, which will be useful in our future work.

We approach the development of this rubric from an iterative design perspective. The current draft of the rubric is in Table 8. Examples of comments that fit each score are in Table 10 in the Appendix.

TABLE VIII
Current Version of Revised Rubric

Feedback Criterion	0	1	2
Contributions to Group Tasks (role fulfillment, task management, task execution, project contributions, work output,)	No task described	General or vague description of how the team member contributes to group tasks, with minimal details.	Specific or detailed description of tasks that the team member contributes to that imply their value to the team
Behavior (interpersonal dynamics, team engagement, communications, personal attributes)	No behaviors described	General or vague description of team member behaviors, with minimal details	Specific or detailed description of team member behaviors that imply their value to the team
Gap (recognition of a negative difference between the ratee's performance and an expected standard)	No gap identified	A gap is alluded to or briefly mentioned, but lacks specific details on how it compares to an expected standard.	A gap is discussed with specific details that highlight or easily imply how the gap compares to an expected standard

Action (suggested steps to remedy gaps or improve performance)	No actions identified	General or vague description of an action with minimal details.	Specific or detailed description of how a team member should address a gap in their performance.

Using the revised rubric, we recoded the data and recalculated the weighted Cohen's kappa for the four criteria (Table 9). The new inter-rater reliability scores show a large improvement with the new rubric. Kappa values from 0.41– 0.60 are considered moderate, from 0.61–0.80 as substantial, and from 0.81–1.00 as almost perfect agreement [28]. The IRR from the second round shows a marked improvement, especially in the Gap criterion, but there is still improvement to be made to distinguish Task from Behavior. The second round gives us confidence that a more reliable and context-appropriate instrument will allow us to evaluate the impact of the feedback intervention with greater efficacy.

TABLE IX
Inter-rater Reliability from Coding Round 2

	Linear K	Quadratic K
Task	0.51	0.61
Behavior	0.30	0.48
Gap	0.83	0.88
Action	0.74	0.82

VIII. Conclusion

This study seeks to understand the impact of an in-class feedback intervention in a first-year engineering PBL course on the quality of peer feedback comments written within two CATME peer evaluations. To adequately evaluate any change in the quality of peer comments between students who participated in the intervention and students who did not, we investigated current theoretical frameworks and rubrics used to evaluate feedback in adjacent fields. Our preliminary data suggests that our intervention, similar to previous forms of feedback training, improved students' peer comments. However, after adopting and slightly modifying a rubric from medical education, we realized several differences in the context of our peer feedback data that limited the rubric's applicability. These differences allowed us to begin a process of iterating the rubric to more effectively assess the quality of peer feedback comments in an engineering PBL context. In future work, we will reanalyze our current data with the new rubric to further validate the rubric and reaffirm this paper's preliminary findings. Additional work is still needed to further examine how improved peer feedback impacts learning and team performance.

IX. References

[1] H. J. Passow, "Which ABET Competencies Do Engineering Graduates Find Most Important in their Work?," *J. Eng. Educ.*, vol. 101, no. 1, pp. 95–118, 2012, doi: 10.1002/j.2168-9830.2012.tb00043.x.

[2] C. L. Dym, A. M. Agogino, O. Eris, D. D. Frey, and L. J. Leifer, "Engineering Design Thinking, Teaching, and Learning," *J. Eng. Educ.*, vol. 94, no. 1, pp. 103–120, 2005, doi: 10.1002/j.2168-9830.2005.tb00832.x.

[3] D. Woods, R. Felder, A. Rugarcia, and J. Stice, "The future of engineering education III. Developing critical skills," *Chem. Eng. Educ.*, vol. 34, Jun. 2000.

[4] M. W. Ohland *et al.*, "The Comprehensive Assessment of Team Member Effectiveness: Development of a Behaviorally Anchored Rating Scale for Self- and Peer Evaluation," *Acad. Manag. Learn. Educ.*, vol. 11, no. 4, pp. 609–630, Dec. 2012, doi: 10.5465/amle.2010.0177.

[5] S. Brutus and M. B. L. Donia, "Improving the Effectiveness of Students in Groups With a Centralized Peer Evaluation System," *Acad. Manag. Learn. Educ.*, vol. 9, no. 4, pp. 652–662, 2010.

[6] S. G. Harkins and J. M. Jackson, "The Role of Evaluation in Eliminating Social Loafing," *Pers. Soc. Psychol. Bull.*, vol. 11, no. 4, pp. 457–465, Dec. 1985, doi: 10.1177/0146167285114011.

[7] C. O. Mayfield and J. R. Tombaugh, "Why peer evaluations in student teams don't tell us what we think they do," *J. Educ. Bus.*, vol. 94, no. 2, pp. 125–138, Feb. 2019, doi: 10.1080/08832323.2018.1503584.

[8] A. Thompson, "Fostering development of teamwork skills in an introductory engineering course," in *2017 IEEE Frontiers in Education Conference (FIE)*, Oct. 2017, pp. 1–4. doi: 10.1109/FIE.2017.8190551.

[9] L. J. Shuman, M. Besterfield-Sacre, and J. McGourty, "The ABET 'Professional Skills' — Can They Be Taught? Can They Be Assessed?," *J. Eng. Educ.*, vol. 94, no. 1, pp. 41–55, 2005, doi: 10.1002/j.2168-9830.2005.tb00828.x.

[10] W. Huang, R. Wynkoop, M. Exter, and F. Berry, "Feedback Matters: Self-and-Peer Assessment Made Better with Instructional Interventions," presented at the 2022 ASEE Annual Conference & Exposition, Aug. 2022. Accessed: Sep. 06, 2023. [Online]. Available: https://peer.asee.org/feedback-matters-self-and-peer-assessment-made-better-with-instructional-interventions

[11] A. C. Loignon, D. J. Woehr, J. S. Thomas, M. L. Loughry, M. W. Ohland, and D. M. Ferguson, "Facilitating Peer Evaluation in Team Contexts: The Impact of Frame-of-Reference Rater Training," *Acad. Manag. Learn. Educ.*, vol. 16, no. 4, pp. 562–578, 2017.

[12] B. E. Mertz, D. M. Ferguson, and M. I. Hoque, "How Competent are Freshman Engineering Students in Constructively Rating Their Peers in a Team Context?," presented at the 2018 ASEE Annual Conference & Exposition, Jun. 2018. Accessed: Sep. 06, 2023. [Online]. Available: https://peer.asee.org/how-competent-are-freshman-engineering-students-in-constructively-rating-their-peers-in-a-team-context

[13] M. Borrego, J. E. Froyd, and T. S. Hall, "Diffusion of Engineering Education Innovations: A Survey of Awareness and Adoption Rates in U.S. Engineering Departments," *J. Eng.*

Educ., vol. 99, no. 3, pp. 185–207, 2010, doi: 10.1002/j.2168-9830.2010.tb01056.x.
[14] Ú. Beagon, D. Niall, and E. Ní Fhloinn, "Problem-based learning: student perceptions of its value in developing professional skills for engineering practice," *Eur. J. Eng. Educ.*, vol. 44, no. 6, pp. 850–865, Nov. 2019, doi: 10.1080/03043797.2018.1536114.
[15] O. Ryan, M. J. Fisher, L. Schibelius, M. V. Huerta, and S. Sajadi, "Using a scenario-based learning approach with instructional technology to teach conflict management to engineering students," presented at the 2023 ASEE Annual Conference & Exposition, Jun. 2023. Accessed: Feb. 06, 2024. [Online]. Available: https://peer.asee.org/using-a-scenario-based-learning-approach-with-instructional-technology-to-teach-conflict-management-to-engineering-students
[16] L. Zhang and Y. Ma, "A study of the impact of project-based learning on student learning effects: a meta-analysis study," *Front. Psychol.*, vol. 14, 2023, Accessed: Feb. 06, 2024. [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1202728
[17] N. Mentzer, A. Jackson, K. A. Richards, A. N. Zissimopoulos, and D. Laux, "Student Perceptions on the Impact of Formative Peer Team Member Effectiveness Evaluation in an Introductory Design Course," presented at the 2015 ASEE Annual Conference & Exposition, Jun. 2015, p. 26.1422.1-26.1422.20. Accessed: Dec. 18, 2023. [Online]. Available: https://peer.asee.org/student-perceptions-on-the-impact-of-formative-peer-team-member-effectiveness-evaluation-in-an-introductory-design-course
[18] A. Burgess *et al.*, "Peer review in team-based learning: influencing feedback literacy," *BMC Med. Educ.*, vol. 21, no. 1, p. 426, Aug. 2021, doi: 10.1186/s12909-021-02821-6.
[19] J. Hattie and H. Timperley, "The Power of Feedback," *Rev. Educ. Res.*, vol. 77, no. 1, pp. 81–112, 2007.
[20] C. Hastie, K. Fahy, and J. Parratt, "The development of a rubric for peer assessment of individual teamwork skills in undergraduate midwifery students," *Women Birth J. Aust. Coll. Midwives*, vol. 27, no. 3, pp. 220–226, Sep. 2014, doi: 10.1016/j.wombi.2014.06.003.
[21] T. Camarata and T. A. Slieman, "Improving Student Feedback Quality: A Simple Model Using Peer Review and Feedback Rubrics," *J. Med. Educ. Curric. Dev.*, vol. 7, p. 2382120520936604, Jan. 2020, doi: 10.1177/2382120520936604.
[22] D. R. Sadler, "Formative assessment and the design of instructional systems," *Instr. Sci.*, vol. 18, no. 2, pp. 119–144, Jun. 1989, doi: 10.1007/BF00117714.
[23] S. Gauthier, R. Cavalcanti, J. Goguen, and M. Sibbald, "Deliberate practice as a framework for evaluating feedback in residency training," *Med. Teach.*, vol. 37, no. 6, pp. 551–557, Jun. 2015, doi: 10.3109/0142159X.2014.956059.
[24] R. M. Abraham and V. S. Singaram, "Using deliberate practice framework to assess the quality of feedback in undergraduate clinical skills training," *BMC Med. Educ.*, vol. 19, no. 1, p. 105, Apr. 2019, doi: 10.1186/s12909-019-1547-5.
[25] A. Field, J. Miles, and Z. Field, *Discovering Statistics Using R*. SAGE, 2012.
[26] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," *Psychol. Bull.*, vol. 70, no. 4, pp. 213–220, 1968, doi: 10.1037/h0026256.
[27] S. Sajadi, O. Ryan, L. Schibelius, and M. Huerta, "WIP: Using Generative AI to Assist in Individual Performance Feedback for Engineering Student Teams," in *2023 IEEE Frontiers in Education Conference (FIE)*, IEEE, 2023, pp. 1–5. Accessed: Feb. 07, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10343517/

[28] M. L. McHugh, “Interrater reliability: the kappa statistic,” *Biochem. Medica*, vol. 22, no. 3, pp. 276–282, 2012.

Appendix

TABLE X
Revised Rubric with Examples

Feedback Criterion	0	1	2
Contributions to Group Tasks (role fulfillment, task management, task execution, project contributions, work output)	**No task described**	**General or vague description of how the team member contributes to group tasks, with minimal to no details.**	**Specific or detailed description of tasks that the team member contributes to that imply their value to the team**
	[name]'s work is satisfactory.	*He helped us with his Solidworks skills and did a good job contributing to our overall goals.*	*[name] was a very open teammate who provided many ideas such as the name for our project's design as well as the temperature sensing leash design which has been chosen to be our main focus on the project. He always got his work done and even offered guidance on anyone else's part of the presentation ...Him showing that he can be a good asset allows us to have full trust in him further on and also know that his work will reflect on his strong effort.*
Behavior (interpersonal dynamics, team engagement, communications, personal attributes)	**No behaviors described**	**General or vague description of team member behaviors, with minimal to no details**	**Specific or detailed description of team member behaviors that imply their value to the team**
	He is a good teammate and does the work that is expected of him.	*He consistently pops out great ideas and is terrific at research and information gathering. Very good at listening and giving feedback.*	*...consistently guides us and maintains a balance between enjoyment and productivity. She creates a welcoming atmosphere, fostering a non-judgmental environment and ensuring everyone feels comfortable sharing their thoughts. Her leadership style seamlessly blends positivity with responsibility, making our sessions both productive and enjoyable.*

Gap (recognition of a negative difference between the ratee's performance and an expected standard)	**No gap identified**	**A gap is alluded to or briefly mentioned, but lacks specific details on how it compares to an expected standard .**	**A gap is discussed with specific details that highlight or easily imply how the gap compares to an expected standard**
	I enjoy working with [name] and I have no issues.	*[name] has missed a few important things in terms of contributing to group work. When he's present and on he's a great contributor but isn't always the easiest to rely on for work.*	*...However, she frequently misses classes (roughly 1/3 of classes attended). Additionally, the only time she reached out to the group chat once before missing a meeting (the 2nd group meeting). This makes the process significantly harder because it is very difficult to get her input when making important decisions during meetings. Since she rarely attends class, in order to complete group work, other group members will have to reach out to her first and tell her which sections to complete. [name] also submits her sections of the work very near the due date which causes a lot of stress and uneasiness among other group members…*
Action (suggested steps to remedy gaps or improve performance)	**No actions identified**	**An action is vaguely described, but lacks any specific details.**	**Specific or detailed description of how a team member should address a defined gap in their performance.**
	Shawn has been a little less of a contributor for this portion of the assignments… due to a significant amount of midterms piling.He has helped a lot with our project's prototyping section.	*He could probably stand to contribute a bit more to the team (not to say he doesn't contribute, just that so far his contributions have not really stood out)... I think he wants to contribute and just has not had a real opportunity yet.*	*...Attending class or notifying the group when she cannot make it would greatly help speed up the process and generate higher quality products. Additionally, it would help to reach out earlier so that group tasks can be reviewed and completed earlier.*