Peer Evaluation Behavior of First Year Engineering (FYE) students and K-12 students

Daniel Ferguson, Yuchen Cao, Mohd Iramul Hoque, Hilda Ibriga, Chad Lally, Matthew Ohland

Purdue University, West Lafayette, USA

dfergus@purdue.edu, cao273@purdue.edu, hoque@purdue.edu, hibriga@purdue.edu, clally@purdue.edu, ohland@purdue.edu

Abstract
A student’s ability to work in a team setting is vital to their college career as well as their work life and therefore it is often a significant factor in a corporate hiring process. Recognizing this need, a number of U.S. undergraduate programs as well as many K-12 instructors, particularly the ones in Project Lead The Way (PLTW) courses use team or project based courses in their curricula. Hundred’s of these programs, including K-12, also use a teamwork formation and assessment system, CATME (Comprehensive Assessment of Team Member Effectiveness), that asks students to assess the teamwork behavior of their peers. Much work has been done in the use of teams in K-12 but not on the analysis of teamwork behavior and the effective use of teams. The goal of our research was to determine whether or not K-12 students behave similarly to First Year Engineering (FYE) students, in particular, when they rate their peers as well as when they are being rated - is the quality of peer evaluations performed by K-12 students similar to the quality of peer evaluations performed by First Year Engineering (FYE) students?

Introduction
According to an article by the Australian Institute of Business, teamwork and the correct team behavior are key attributes sought after by a large number of companies when hiring new employees. In addition to that, they state that working in teams not only help distribute the workload better but lead to greater efficiency, better communication in the future as well as creates a supportive environment for workers that can serve as a platform for even better performance. [1] Hence, teamwork skills training has become more prevalent throughout college programs and in businesses [2]. Teamwork is defined as “a cooperative or coordinated effort on the part of a group of persons acting together” [3]. Chen argues that many students entering the workplace lack key teamwork skills that hamper their abilities to excel in their job field [4]. Teamwork is not only deemed important in higher education, but has also began to attract the widespread attention of a multitude K-12 instructors in recent times [5], especially those teaching Project Lead The Way (PLTW) courses. Many K-12 instructors throughout the nation have started using small teams in their courses similar to that done in most college institutions. Although Wang and MacCann, etc. used self-report, other-report and situational judgment test to assess high-school students’ teamwork skills, their work simply adopted the measures that were perhaps not as well-designed as CATME, a tool that allows you to capture behavior on a number of different dimensions. In addition to that, they’ve also failed to include a longitudinal study to understand the trends/shifts in a
student's behavior over time [5]. The development of a systematic approach of assessing the quality of teamwork is critical for K-12 classrooms just like in higher education. Thus, adopting CATME as an assessment tool for use in evaluating the effective use of teams in K-12 promises a lot of potential, given that there is a large similarity in the ratings and the quality of ratings between K-12 and FYE students,

Our primary research goal was to determine how similarly High School students and FYE students rated themselves and their peers when using CATME peer evaluations to help themselves and their teammates learn and improve their abilities when it comes to working in teams. A second research question was: is there a significant difference in the quality of peer ratings across dimensions between these two groups of students? The quality of a rating, in this paper, is defined as having a larger dispersion (measured by standard deviation) in the ratings of teamwork dimensions or in the ratings of team members on specific teamwork dimensions. It is our understanding that a larger dispersion shows that an individual is correctly differentiating their perception of a teammates contribution on each particular dimension. While it is clearly possible that an individual exhibits the same level of performance across all 5 CATME dimensions, it is not likely, and similarly it is unlikely that there are no differences in teamwork performance among all team members on any of the 5 CATME dimensions [6, 7].

CATME (Comprehensive Assessment of Team Member Effectiveness) is a web based tool created for the academic teamwork environment and is used to assist students in giving peer feedback to their team members [8]. Prior research has shown that the accountability of ratings given to or by an individual play a huge role in reflecting the accuracy of the rating and this feature is an integral part of CATME [9,10]. In addition to that, CATME makes use of a number of analytical models such as Dispersion Analysis, Convergence Analysis and the Social Relations Model (SRM) [11]. This paper discusses the results obtained from the execution of the Dispersion Analysis on data comparing ratings produced by K-12 students and FYE students.

CATME is constructed around five behavioral dimensions: Having Relevant Knowledge, Skills and Attributes (KSAs), Contributing to the Team’s Work, Interacting With Teammates, Keeping the Team on Track, and Expecting Quality [12, 13]. These dimensions are defined as follows:

**Having (H)** relevant KSAs refers to the base knowledge of individual team members. It means having the required skills to solve the problems at hand, or being willing to learn the skills an individual lacks.

**Contributing (C)** to the Team’s Work is being able to add value to your team’s work/project. It includes completing your portion of the work in a timely fashion.

**Interacting (I)** with teammates refers to the way individuals communicate within their teams. Encouraging every team member to give their opinion and allowing their voice to be heard and respecting their ideas is the essence of good interacting.

**Keeping (K)** the team on track is similar to being a timekeeper. When an individual is aware of the timeline for the project and makes sure their team meets the required steps in time, they are helping the team stay on track.
Expecting (E) quality is taking expectations to the next level and working collaboratively to produce the best possible team outcomes. Every aspect of the five teamwork dimensions is equally important to team success and a critical element in the peer reviews [12].

Peer evaluations facilitate better learning outcomes in upper level education, encouraging students to continue their engagement with constructive team behavior in future team activities [14]. A peer evaluation is an assessment of an individual’s contribution to a work activity by their peers, either as students, or as professionals in industry [15]. Peer evaluations help to teach individuals and teams how to act and how to assess one another’s performance. Peer evaluations, as facilitated by the CATME system, potentially point out behaviors where an individual excels and areas where he/she may need improvement [16, 17].

Research Methods
I. CATME Peer Evaluation
CATME peer reviews are conducted online and students receive peer review feedback in an online form. Peer review feedback is shown as pointers to word descriptions of behaviors similar to their own rating or better or worse than they were rated on average by their peers as shown in Fig 1. No numbers are provided to students and behavior descriptions appropriate to their average ratings are cited. For each of the five CATME peer rating dimensions students see their own ratings of themselves, their average ratings by their teammates and the average ratings for all team members as shown in Fig 1. Students see only average ratings interpreted as pointer placements on feedback screens and do not see their actual numeric ratings by their teammates as all ratings are held confidential, although numeric rating data is provided to instructors. [18]

Figure 1: Feedback Screen for Contributing Dimension of the CATME Peer Review Tool
II. Data Collection

CATME peer evaluations were completed by 2 groups of K-12 students enrolled in a Physics class, taught by the same instructor (9th grade and 12th grade), at a college prep high school in the Pacific Northwest. The 9th grade class was composed of 11 teams of 3 and the 12th grade class was composed of 10 teams of 3. Both the 9th grade and 12th grade class were split roughly 50-50 in terms of gender and 30% of the 9th grade class while 26% of the 12th grade class were composed of minorities (Asian, Black, Hispanic, or Other). 9th graders rated themselves and their peers on 3 dimensions (C, I & E) and 12th graders did so on 4 dimensions (C, I, K & E). This was done based on the discretion of the instructor and it is our assumption that he/she only felt the need to measure 3 dimensions for the 9th graders but felt that an additional dimension was important for the 12th graders. Students were placed in teams using CATME’s Team Maker tool and went on to complete a semester long project (in the Fall of 2016) throughout which they took part in 3 separate peer evaluations providing both quantitative and qualitative feedback. For the purpose of our research, only the quantitative data was analyzed using the Dispersion Analysis tool developed by CATME.

CATME peer evaluations were also completed by First-Year Engineering (FYE) students at a major Midwestern university in the Fall of 2015. FYE students at this institution number over 1,600 (annually) including 23.1% women and 6.4% minorities. Each FYE section was composed of up to 120 students and there were 16 total sections. Although, for the purpose of our research only the teams of 3 from this data set was used to compare to data obtained from K-12 students. This was done in order to more closely match the number of data points that were used. There were 6 teams of 3 in the Fall 2015 FYE data set that was compared to K-12 data. In addition to that, despite having data for all 5 dimensions, only 3 dimensions (C, I & E) were used when comparing to the 9th grade data set and 4 (C, I, K & E) when comparing to the 12th grade data set to maintain consistency.

III. Data Analysis

The measure of dispersion used in this analysis is defined as the standard deviation of each students’ rating of themselves as well as their teammates across the 3 (or 4) dimensions being used. To be precise, Figure 2 shows a sample of the raw quantitative peer evaluation data. For a K-12 team with 3 members completing a peer evaluation on 4 dimensions (C, I, K & E), each rater in the team contributes to a 3×4 rating matrix; with all 3 raters’ response combined column wise forms a 3×12 matrix. The standard deviations for each student were calculated for each row. In Figure 2, this was the standard deviation for each of the three row under the four “Rater 1” columns. Then the three row-wise standard deviations were averaged, then placed in a matrix of average dispersions and referred to as the dispersion for the Rater 1’s ratings for all the team members including Rater 1. This procedure was repeated for Rater 2 and 3 accordingly. The same methods were used to calculate the dispersion matrix for the rest of the K-12 teams as well as the FYE control group with the same dimensions (C, I, K & E in this example). For reviews two and three, the same calculation were repeated.
A One-Way ANOVA was used to compare the differences in dispersions between K-12 and FYE control samples for each of the three peer reviews. The time gaps between reviews are around 4 weeks for both the K-12 and FYE control group. The difference in dispersion is then calculated using SAS where the dispersion of ratings between two different reviews are compared. Holm-Bonferroni step down adjustment for multiple comparisons was adopted to control family-wise error so that we could obtain more powerful results compared with the original Bonferroni method [19]. A statistically significant P-value (with Holm-Bonferroni step down adjustment) from this One-Way ANOVA test (less than 0.05) tells us that the difference in dispersion and/or average ratings between reviews is substantial, meaning that there is reason to believe that the two samples behave differently in terms of how individuals rated each other, otherwise the difference is believed to have stemmed from random error. The effect sizes Cohen’s d are calculated using differences in estimates and the pooled standard deviations of the deviation factors. According to Cohen’s guidelines, an effect size d around 0.2 is considered small, 0.5 is considered medium and 0.8 is considered as large [20].

<table>
<thead>
<tr>
<th>Student ID</th>
<th>Team ID</th>
<th>Rater #</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>I</td>
<td>K</td>
</tr>
<tr>
<td>A001</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>A002</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>A003</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>
Findings
Table 1 - Dispersion Analysis Results

<table>
<thead>
<tr>
<th>Control Sample</th>
<th>K12 Sample</th>
<th>Peer Review Time</th>
<th>Difference in Standard Deviation</th>
<th>F-statistic</th>
<th>Holm-Bonferroni Adj. P_value of Difference in Standard deviation</th>
<th>Effect size (Cohen’s d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FYE FALL 15</td>
<td>K12-9th</td>
<td>1</td>
<td>-0.1002</td>
<td>0.185</td>
<td>0.6705</td>
<td>0.138</td>
</tr>
<tr>
<td>FYE FALL 15</td>
<td>K12-9th</td>
<td>2</td>
<td>0.1363</td>
<td>3.349</td>
<td>0.1456</td>
<td>0.588</td>
</tr>
<tr>
<td>FYE FALL 15</td>
<td>K12-9th</td>
<td>3</td>
<td>0.1762</td>
<td>5.382</td>
<td>0.0672</td>
<td>0.733</td>
</tr>
<tr>
<td>FYE FALL 15</td>
<td>K12-12th</td>
<td>1</td>
<td>0.2455</td>
<td>0.865</td>
<td>0.5614</td>
<td>0.320</td>
</tr>
<tr>
<td>FYE FALL 15</td>
<td>K12-12th</td>
<td>2</td>
<td>0.1625</td>
<td>3.312</td>
<td>0.2337</td>
<td>0.622</td>
</tr>
<tr>
<td>FYE FALL 15</td>
<td>K12-12th</td>
<td>3</td>
<td>0.1201</td>
<td>1.210</td>
<td>0.5614</td>
<td>0.351</td>
</tr>
</tbody>
</table>

Each K-12 class was compared with the FYE Fall 15 section throughout the three reviews. The table above shows the results from the One-Way ANOVA test in the dispersion analysis. Peer reviews of the same time period were compared. The difference in dispersion for both comparisons across all reviews was statistically identical. Hence, this has led us to believe that the way students in K-12 teams rate each other are very similar to the way college students rate one another when working in teams.

As for the quality of the ratings, as defined earlier the higher the dispersion the better we perceive the given ratings as. In the case of the dispersion analysis, the difference in dispersion was calculated by subtracting the dispersion seen in the FYE Fall 15 sample from the K-12 sample. A positive difference in dispersion actually means that the dispersion in the K-12 sample was larger than that seen in the FYE Fall 15 section which actually suggests that the quality of ratings produced by K-12 students is better than those produced by FYE students.

The effect sizes for most of the comparisons are medium according to Cohen’s guidelines [20], except for that of the comparisons for review 1, which means the experiments as a whole have medium effects in differentiating the differences between the two samples.
Conclusion
From this analysis we have concluded that it is highly likely that K-12 students in teams in college preparation schools behave similarly to FYE students placed in teams in regards to rating themselves and their peers using CATME. Extending the use of CATME to more K-12 institutions will not only encourage more instructors to incorporate teamwork peer evaluations in K-12 education but also help K12 teachers to critically evaluate the performance of their students’ teamwork behavior.

Further Research
This analysis is being extended using two other analytical models, namely, the Convergence Analysis and the Social Relations Model (SRM) [21]. Convergence Analysis involves assessing changes in the difference between self ratings and the ratings other team members give that individual in each review. The results of each review are compared over time to see how similar or dissimilar self ratings and peer ratings become. SRM on the other hand assesses the changes in different variance components (rater, target and relationship) over time. The rater variance measures the variance in ratings given to peers by an individual, target variance measures the variance in ratings received by an individual and lastly the relationship variance assesses the dyadic variance in ratings for each pair of individuals in the team. There are numerous college institutions that are making use of CATME for research and non-research purposes and we hope to extend this research to K-12 students as well.

Limitations
This pilot experiment was done with limited samples from one K-12 school. Larger samples from more K-12 schools may help get a better understanding of K12 student performance using CATME peer evaluations. Furthermore, with a larger sample size we are able to credibly extend our analysis to other analytical models such as the Convergence Analysis Model and the Social Relations Model (SRM).
References