Gendered Influences on Peer Assessments of Team Behavior

Emily Miller
Department of Systems and Information Engineering
University of Virginia
Charlottesville, VA, USA
emsumi94@gmail.com

Reid Bailey
Department of Systems and Information Engineering
University of Virginia
Charlottesville, VA, USA
rrbailey@virginia.edu

***Abstract*— This Research Full Paper will examine how gender may correlate with peer perceptions of behaviors on engineering student project teams. While peer assessments are often used to encourage personal responsibility for group tasks (and discourage "social loafing"), they may fail to account for the way students' attributes contribute to their perceptions of the behavior of peers. These perceptions may be inconsistently held from one person to the next and are driven by various factors (like gender), including the attributes of the judger, the individual being judged, the relationship between the two, and/or the group at large. This study seeks to answer: Does the perception of a peer's team behavior (in terms of frequency or severity) vary with the gender identity of the individuals involved?**

This study uses a newly-created survey called the Team Behaviors and Attitudes Survey to assess the subjective perceptions of each of eleven undesirable teammate behaviors. Data has been collected from 47 teams, involving 329 unique pairings of students. For each of these behaviors, proportion tests and chi-square hypothesis tests are used to assess whether gender identity affects how severely and frequently behaviors are perceived, respectively. Males were found to often rate others more severely than females, regardless of the gender of the target. Other results were mixed, emphasizing the particularities inherent in perception creation.

Keywords—teamwork, gender

I. INTRODUCTION

Group work is an essential component of engineering classrooms. Studies have shown that working in small groups promotes greater academic achievement, more favorable attitudes toward learning, and increased persistence [1]. Group work is also thought to prepare students for the "real world" in engineering industry, where teamwork is prevalent. Graduates of engineering programs have reported that learning how to work in multidisciplinary teams was a key aspect of their post-graduate industry positions, spending 60-80% of their time working on a team with other engineers [2]. A key reason for this is that working on teams allows larger engineering challenges to be tackled than one engineer alone could have completed. Instructors may also want students to learn from one another – one student may have strengths, knowledge, or skills another does not – and thus the quality of the final product can be enhanced.

Teams, however, are made of individuals that are imperfect – thus, the conditions of teams may not always be ideal. How students experience their teams is certainly of importance for any instructor wanting to support student persistence in engineering [3]-[5]. While some students perform at high levels on teams, other students may fail to contribute to the extent their instructor or teammates desire.

Understanding the subjective student experience in project teams is important in terms of helping students learn how to be better teammates in school and the workplace, designing more relevant peer assessments, and ultimately positively impacting students' motivation and persistence in engineering. Perceptions of teammate behavior can be driven by other factors than simply observance of behavior. Social identity theory [6] posits that easily-accessible differences between people, like gender, inform social categorization of people into groups, social identification of the groups to which we ourselves belong, and social comparison, in which in-group/out-group distinctions frame our perceptions of our own group and others'. Thus, gender may be one factor that helps inform our discernment of how others behave.

II. LITERATURE REVIEW

A. Behaviors on Teams

Team environments in educational settings are unique in that the performance of a team (and thus, of an individual) is impacted by the behavior and idiosyncrasies of different team members. These behaviors may aggregate by way of composition, in which all team contributions are weighted equally and the team outcome is a linear combination of team member inputs, or compilation, in which more complex relationships emerge [7]. Either way, team outcomes depend on the contributions of its members. [7], [8] For an individual participant on a team that is not functioning well, this means that one's grade, ability to attain course learning objectives, and perspective on their own learning experience may be negatively impacted by the failure of other team members to contribute effectively.

To mitigate this risk, many educators have adopted peer assessments. These assessments typically ask students to rate one another's performance on a Likert scale [9], [10]. When adopted at intervals during the project, these assessments can be used formatively to help diagnose problems and correct errant behaviors. They can also be used as a motivational tool to encourage active participation by using teammate perceptions of an individual as a factor in one's project grade [11] Teamwork often takes place away from the eyes of the instructor; thus, these tools encourage accountability and discourage "social loafing" – a phenomenon in which individuals tend to exert less effort when working collectively than individually. [12] Furthermore, these peer assessments have been found to promote cooperation, higher levels of performance, and team member satisfaction. [13]

A taxonomy of behaviors typically evaluated by instructors in these assessments was developed by Baker in a meta-analysis of the literature on this subject. She identifies eight components (Table 1) [14].

As these evaluation tools are developed by instructors, the desired behaviors as listed are top-down rather than bottom-up. How the students themselves are perceiving their own learning environment is vitally important to their persistence in engineering [3]-[5]. In prior work, eleven behaviors have been identified by peers as undesirable in group settings. [15] These include expecting too much from others, failure to advance the project toward completion, failure to prioritize the project, inconsistency of contributions, inconsistency with an engineering identity, lack of communication, lack of competence/experience/skills, lack of initiative, procrastination, restricting others' work, and unreliability. Though many of these undesirable behaviors overlap with ones identified by instructors, some aspects of teammate behavior viewed as important to students are not reflected in most instructor-created peer assessments. The mapping of Baker's meta-analysis and these eleven undesirable behaviors can be found in Table 1.

Table 1: Comparisons between positive behavioral components and negative emergent categories

Meta-analysis of Effective Team Behavioral Components (Top Down) [14], [15]	**Emergent Categories of Undesirable Teammate Behaviors (Bottom-up)** [15]
Attended group meetings; was available and on time	Failing to prioritize project
Submitted quality work	Lack of competence, experience, or skills
Exerted effort and took an active role	Failing to advance toward project's completion; Lack of initiative
Cooperated and communicated with others	Lack of communication
Managed group conflict	N/A
Made cognitive contributions; possessed and applied necessary knowledge and skills	Lack of competence, experience, or skills
Provided structure for goal achievement	N/A
Was dependable, kept his or her word	Unreliability, Procrastination, Inconsistency of contribution
N/A	Expecting too much from others
N/A	Inconsistency with an engineering identity
N/A	Restricting the work of others

The components (in terms of undesirable behaviors) that emerged from the bottom-up, student perspectives which were not in the top-down instructor view include expecting teammates to contribute beyond their "fair share", possessing traits (personality, motivation, etc.) that seem to conflict with an engineering identity, and directly or indirectly inhibiting the group from completing its work in a timely manner [15]. The two components that were only in the top-down instructor view were both specific management skills related to running a successful team.

B. Expertise Recognition

Thomas-Hunt and Phillips note that perceptions do not always match reality by saying, "the complexity of most organizational tasks makes it difficult for expert members to demonstrate the correctness of their perspective prior to the completion of the group's task and the receipt of feedback from sources external to the group. Consequently, teams often have difficulty assessing the veracity of members' claims of expertise" [16]. Evaluations of expertise within a group may differ between individuals, and thus may be colored by the idiosyncrasies of those individuals. These evaluations also may or may not reflect the true nature of the object of evaluation. Some studies have found that among women, educational status as objectively measured by education level attained does not significantly predict expertise evaluations [16], [17]

This may be explained in part by social identity theory [6] which posits that easily-accessible differences between people, like gender, form the basis for in-group/out-group distinctions. People tend to prefer the in-group in a context-specific way. For example, a male job applicant may be preferred for a traditionally "male" position such as police chief whereas the female job applicant may be preferred for a traditionally "female" position such as nurse. These preferences are not always explicitly known to the evaluator. One study found that male and female applicants for the same position were not found to actually possess different strengths, but rather that the evaluator redefined the criteria deemed necessary for job success to favor the specific credentials that the in-group candidate happened to possess. The evaluators in this case were not aware of their own implicit biases. In fact, perceiving one's own judgments as objective actually predicted greater bias [18] As engineering is widely known to be a male-dominated field, women are often perceived to be members of the "out-group" in this context. Thus, expertise judgments made about peers may be influenced by the implicit biases associated with the in-group/out-group distinction.

The greatest variance in expertise evaluations has been shown to be exhibited not through the attributes (such as gender or education level completed) of the person being evaluated (i.e., the target) as one might expect, but rather through the attributes of the person providing the evaluation (i.e., the actor) and the relationship between the two individuals (i.e., the dyad) [17]. Instructors may hope that high-performing students will help weaker students on a diversely-performing project team. However, Van Der Vegt suggests that levels of interpersonal commitment to help one another felt between individuals on a team were predicted by the relationship between those two group members. High-performing students were apt to help other high-performing students, leaving students with weaker histories of performance to rely on other weak students. [19].

III. METHODS

The Team Behaviors and Attitudes Survey (TBAS) explores teammate perceptions of one another's behavior over the course of the class project in which they participated using a round-robin format. In particular, it examines eleven distinct types of undesirable teaming behaviors identified in previous work [15], including expecting too much from others, failing to advance toward the project's completion, and nine others detailed in Table 1. Though tools such as CATME Peer Evaluation [20] have been utilized by other researchers to quantify peer perceptions of one another, this work is focused only on the negative aspects of perception through examination of these eleven identified undesirable behaviors in particular. Dividing the perceptions into these eleven behaviors also allowed a level of granularity that the CATME system did not support. The survey also collected demographic information, such as sex, age, and year in school to analyze whether perceptions of contribution differed between groups.

The final survey was built in a branched format. First, respondents answered demographic questions about themselves such as their year in school, major, and age. They also indicated their team's size and their teammates' first names (Appendix C). They answered a series of 3 questions about their overall impression of contributions made to the project by each teammate. Then, they read a list of 11 vignettes that corresponded to the 11 identified negative team behavior types [15]. To reduce fatigue and order effects, all 11 behavior vignettes were presented at once. Listing previously named teammates one at a time, participants answered three quantitative questions about their overall viewpoint about each teammate's contributions. Then, respondents indicated whether each teammate exhibited each behavior binarily (yes/no) in one long form. Participants were asked to be particularly sensitive, answering "yes" if a behavior occurred even once. Answering "no" constituted a score of "0" for that particular behavior and teammate. Answering "yes" led to a series of four quantitative and three qualitative questions about that teammate and behavior. The four quantitative scores assessed frequency, effect on personal experience, effect on team's work quality, and effect on team overall. These last three question scores were averaged, yielding a score between 1 and 7 for that teammate. A score of "4" indicated "no effect". Scores above four indicated increasingly positive evaluations of the behavior while scores below four indicated increasingly negative evaluations. In total, behaviors were scored from 1 to 7, so that each teammate had a score for each behavior (Table 2). This branched format provided multiple benefits. First, it allowed participants to compare the behavior vignettes to one another and see the differences more clearly. Second, if respondents knew that answering "yes" required more work beforehand, they might have been less honest about which behaviors did and did not occur.

Table 2: Sample collected data

	Demographics				Team Member #1					Team Member #2				
	Sex	Discipline	Age	...	Overall	Behavior 1	Behavior 2	Behavior 3	...	Overall	Behavior 1	Behavior 2	Behavior 3	...
Participant 1	F	Systems	19		3.7	0	0	0		5.1	5.21	0	6.27	
Participant 2	M	Computer	21		4.2	0	3.95	0		4.4	5.33	2.87	0	
Participant 3	M	Civil	22		2.8	4.26	5.13	0		4.3	0	0	4.33	

Table 3 highlights the relative proportions of male and female participants by race, year in school, and major. In groups with higher n values (white students, fourth years, and students from biomedical or systems engineering), the percentage of students who are male and female are close to 50%. This gender parity was specific to the study, as the engineering school at the University of Virginia is 34% female as a whole. [21] Furthermore, the classes recruited for this study were specifically project design classes. The structure of students in each school year is a result of the curriculum. Most students do not take classes with long-term project components until their third or fourth year. ENGR 1420 is one of the few project classes offered to first years and is open only to a specific group of students from a scholar program.

Table 3: Participant demographic counts by gender

	n	Female	Male
OVERALL	89	50.6%	49.4%
RACE			
Other	3	66.7%	33.3%
Black	3	100%	0%
Asian	15	53.3%	46.7%
White	68	48.5%	51.5%
YEAR IN SCHOOL			
4+	3	100%	0%
Fourth year	60	45%	55%
Third year	20	65%	35%
Second year	0	0%	0%
First year	6	50%	50%
MAJOR			
Aerospace Engineering	4	25%	75%
Biomedical Engineering	37	48.6%	51.4%
Civil Engineering	6	100%	0%
Computer Engineering	3	33.3%	66.7%
Engineering Science	3	33.3%	66.7%
Mechanical Engineering	8	37.5%	62.5%
Systems Engineering	28	53.6%	46.4%

This dataset provided behavior information on 178 unique “ratees”, or "targets" some of whom were also raters, or "actors." 61% of unique targets were male, reflecting team makeup more than anything else. This yielded 329 unique “pairings”, or “dyads”. Out of these 329 dyads, each behavior existed with the following frequencies (Table 4):

Table 4: Behavior frequencies by dyads

Expecting too much from others	13.4% (44 dyads)
Failure to advance project toward completion	13.7% (45)
Failure to prioritize project	16.4% (54)
Inconsistency of contribution	17.0% (56)
Inconsistency with an engineering identity	7.0% (23)
Lack of communication	13.7% (45)
Lack of competence, experience, or skills	9.4% (31)
Lack of initiative	19.1% (63)
Procrastination	9.4% (31)
Restricting others’ work	6.4% (21)
Unreliability	9.4% (31)

The TBAS asks participants to provide demographic information about themselves but not about their teammates (other than first names). To reconstruct likely genders for targets, the following procedure was used. First, some targets were also actors and thus had a listed gender. Second, qualitative data provided about a target sometimes used gender-specific pronouns, which were taken to be the correct gender. Third, the 1996 census list (average respondent age was 21.07 and the survey was taken in 2017) of most popular boys and girls names was searched and the higher listing was chosen as the correct gender. Fourth, if the name was not on the census list (which were most often international names) a google-search was completed of the gender commonly associated with that name. In total, 93 (46%) were previously indicated as an actor, 32 (16%) were taken from qualitative data pronouns, 66 (33%) were taken from the 1996 census list, and 10 (5%) from a google search. The four dyads (Male rates male, Male rates female, Female rates male, and Female rates female) were thus constructed by comparing the actor’s gender to the target’s gender.

IV. RESULTS

In both cases (frequency and severity) Bonferroni corrections were not applied due to the large number of groups – the adjusted alpha would be too conservative. Furthermore, p-values, particularly for frequency, were considered to be sufficiently low to ease concerns of Type I error.

A. Frequency of Perceptions

Frequency of ratings was examined by considering each behavior separately in the Full Pairings dataset. Each participant had different numbers of teammates and thus may have greater or fewer opportunities for rating a behavior as occurring or not. Frequencies were calculated by first constructing proportion charts for the four dyads, as seen for *Inconsistency of contributions* in Table 5. In this case, there were 114 opportunities for male participants to rate male teammates when all unique pairings are considered. Of those, 9 instances of *Inconsistency* were perceived to occur, yielding a proportion of 0.079.

Table 5: Example of frequency proportion chart for Inconsistency of contributions

	“Yes” count	“No” count	Total count	“Yes” proportion
Male-Male	9	105	114	0.079
Male-Female	10	39	49	0.204
Female-Male	23	63	86	0.267
Female-Female	14	66	80	0.175

Chi square tests were conducted to test for differences between frequencies of each of the four dyads (Male-male, male-female, female-male, and female-female). Significant differences were detected in only a minority of behaviors: *Inconsistency of contributions* and *Lack of initiative.* The results for the chi-square tests are shown in Figure 4 below; "m-m" corresponds to males rating males, "m-f" to males rating females, and so on. In both cases, the proportion of females rating males with these behaviors was high. For *Inconsistency of Contribution*, the largest contributors to the chi square are the "male-male" and "female-male" dyads (i.e. whenever there was a male target). For *Lack of initiative*, the largest contributors to the chi square were the "female-male" and "female-female" dyads (i.e. whenever there was a female actor).

(a)

Rows: Dyad Columns: Behavior

	yes IC	no IC	All
m-m	9 19.40 5.5786	105 94.60 1.1443	114
m-f	10 8.34 0.3302	39 40.66 0.0677	49
f-m	23 14.64 4.7764	63 71.36 0.9798	86
f-f	14 13.62 0.0108	66 66.38 0.0022	80
All	56	273	329

Cell Contents
Count
Expected count
Contribution to Chi-square

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	12.890	3	0.005
Likelihood Ratio	13.558	3	0.004

(b)

Rows: Dyad Columns: Behavior

	yes LI	no LI	All
m-m	18 21.83 0.6719	96 92.17 0.1591	114
m-f	13 9.38 1.3943	36 39.62 0.3302	49
f-m	24 16.47 3.4448	62 69.53 0.8159	86
f-f	8 15.32 3.4969	72 64.68 0.8282	80
All	63	266	329

Cell Contents
Count
Expected count
Contribution to Chi-square

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	11.141	3	0.011
Likelihood Ratio	11.360	3	0.010

Figure 1: Chi-Square test for association results for Inconsistency of Contribution (a) and Lack of Initiative (b)

To investigate whether differences in frequency correlate with the actor's gender or target's gender, the datasets above were combined. For example, for *Inconsistency of contributions* there are 19 ratings with a male actor out of a possible total 163. The only statistical differences in terms of the proportions of frequency for male and female actors were for two behaviors: *Inconsistency of contributions* and *Failure to prioritize project.* The results for these tests of proportions are detailed in Table 6, 7, and 8.

Table 6: 2-sample test of proportions results for actors by behavior

Behavior	**Male actors (n=163)**	**Female actors (n=166)**	**Actor z-value**	**Actor p-value**
Expecting too much from others	13.5%	13.3%	0.06	0.95
Failure to advance project toward completion	14.1%	13.3%	0.23	0.82
Failure to prioritize project	20.8%	12.0%	2.17	**0.03***
Inconsistency of contribution	11.6%	22.3%	-2.60	**0.01***
Inconsistency with an engineering identity	6.7%	7.2%	-0.17	0.864
Lack of communication	11.0%	16.3%	-1.38	0.17
Lack of competence, experience, or skills	8.6%	10.2%	-0.51	0.61
Lack of initiative	19.0%	19.3%	-0.06	0.95
Procrastination	11.7%	7.2%	1.38	0.17
Restricting others' work	6.7%	6.0%	0.27	0.79
Unreliability	11.0%	7.8%	1.00	0.319

Table 7: 2-sample test of proportions results for targets by behavior

Behavior	Male targets (n=200)	Female targets (n=129)	Target z-value	Target p-value
Expecting too much from others	15.0%	10.9%	1.11	0.27
Failure to advance project toward completion	13.5%	14.0%	-0.12	0.91
Failure to prioritize project	16.0%	17.1%	-0.25	0.80
Inconsistency of contribution	16.0%	18.6%	-0.61	0.54
Inconsistency with an engineering identity	5.5%	9.3%	-1.26	0.21
Lack of communication	13.5%	14.0%	-0.12	0.91
Lack of competence, experience, or skills	10.5%	7.8%	0.86	0.39
Lack of initiative	21.0%	16.3%	1.09	0.28
Procrastination	11.0%	7.0%	1.28	0.20
Restricting others' work	7.0%	5.4%	0.59	0.56
Unreliability	12.0%	5.4%	2.16	**0.03***

Furthermore, 2-sample tests of proportions were employed to test for differences in the proportions between male and female targets. Only *Unreliability* shows a significant difference. Frequency of judgments about particular team behaviors do not depend on the gender identity of the dyad, actor, or target except in specific cases.

B. Severity of Perceptions

Multifactor ANOVAs with behavior severity as the response and both actor gender and target gender as independent variables were run using the Full Pairings dataset to examine this question. Each of the eleven behaviors was examined separately for normal distribution, homoscedasticity, and independence (see Table 8). In each case, only pairings in which a particular behavior was said to exist were considered, thereby eliminating all of the severity scores of zero associated with the behavior not being present. Equal variances are reported using multiple comparisons.

Table 8: Severity assumption checks

		Equal variance test by gender		
Behavior	**N (% Female)**	**Actor**	**Target**	**AD p-value**
Expecting too much from others	41 (50%)	Equal	Equal	Non-normal, p=0.03
Failure to advance project toward completion	42 (48.9%)	Equal	Equal	Normal, p=0.48
Failure to prioritize project	53 (37%)	Equal	Equal	Normal, p=0.54
Inconsistency of contribution	49 (66.1%)	Equal	Equal	Normal, p=0.52
Inconsistency with an engineering identity	22 (52.2%)	Equal	Equal	Normal, p=0.06
Lack of communication	41 (60%)	Equal	Equal	Normal, p=0.76
Lack of competence, experience, or skills	29 (54.8%)	Equal	Equal	Normal, p=0.65
Lack of initiative	62 (50.8%)	Equal	Equal	Normal, p=0.24
Procrastination	30 (38.7%)	Not equal, p=0.02	Equal	Normal, p=0.54
Restricting others' work	21 (47.6%)	Equal	Equal	Normal, p=0.25
Unreliability	31 (41.9%)	Equal	Equal	Normal, p=0.72

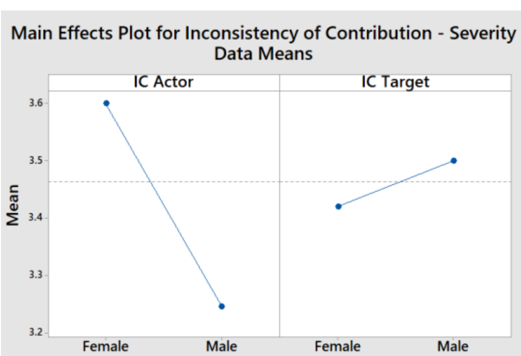
Eleven multifactor ANOVA tests were conducted using a model with actor gender, target gender, and actor gender*target gender as predictors and the specified behavior severity scores as the response. The interaction term is a measure of the impact of the four dyads: Male rates male, male rates female, female rates male, and female rates female. For instance, if males rate males differently than they rate females, that would show up in the interaction term. Results are detailed in Table 9.

Table 9: Multifactor ANOVA significance results

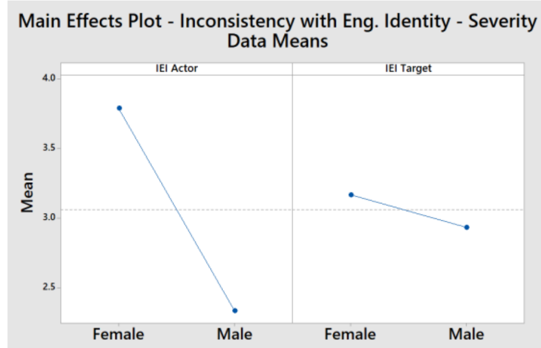
Behavior	Actor gender p-value (Difference in means)	Target gender p-value (Difference in means)	Actor* Target p-value
Expecting too much from others	0.56 (+0.216)	0.59 (+0.164)	0.56
Failure to advance project toward completion	**0.03(*)** (+0.616)	0.53 (+0.167)	0.28
Failure to prioritize project	0.40 (+0.222)	**0.05(*)** (+0.481)	0.74
Inconsistency of contribution	**0.05(*)** (+0.354)	0.65 (-0.08)	0.39
Inconsistency with engineering identity	**<0.01(*)** (+1.455)	0.59 (+0.234)	0.76
Lack of communication	**0.02(*)** (+0.755)	0.20 (-0.237)	0.94
Lack of competence, experience, or skills	0.38 (+0.378)	0.95 (+0.007)	0.71
Lack of initiative	0.83 (+0.145	**0.03(*)** (-0.417)	0.54
Procrastination	**0.03(*)** (+0.759)	0.68 (+0.201)	0.66
Restricting others' work	0.65 (+0.300)	0.43 (+0.428)	0.90
Unreliability	0.18 (+0.427)	0.72 (+0.117)	0.89

(*) indicates significance, α=0.05

( ) shows difference in means (Female – Male). Positive values in difference column mean that males are more severe in their ratings



(a)



(b)

Figure 2: Main effects plots for *Inconsistency of contribution* (a) and *Inconsistency with an engineering identity* (b)

For behaviors, scores range from 1 to 7. A score of "4" indicates "no effect". Lower scores indicate a more negative or severe appraisal of a target's behavior while higher scores indicate progressively positive appraisals of identified behaviors. But which group rates others more severely? For every one of the eleven behaviors, women rated others less severely than men regardless of the target's gender, albeit to different extents. This effect was significant at α=0.05 for *Failure to advance project toward completion* (Female M=3.40, Male M=2.79), *Inconsistency of contribution* (Female M=3.60, Male M=3.25), *Inconsistency with an engineering identity* (Female M=3.79, Male M=2.33), *Lack of communication* (Female M=3.40, Male M=2.65), and *Procrastination* (Female M=3.30, Male M=2.54). Figure 2 shows this effect for two of these behaviors: *Inconsistency of contribution* and *Inconsistency with an engineering identity.*

The target's gender rather than the actor's gender is significant for two behaviors: *Failure to prioritize project* and *Lack of initiative.* Women are rated less severely for *Failure to prioritize the project,* while being rated more severely for demonstrating *Lack of initiative.*

Multifactor ANOVAs indicate that the actor gender often has a more significant impact on how severely a target is rated than the target gender or gender identity of the dyad. For no behavior is the interaction between actor and target genders significant.

Table 10: Summary of significant results for frequency and severity of undesirable behaviors

	Actor*		**Target****	
	Frequency	Severity	Frequency	Severity
Expecting too much from others				
Failure to advance project toward completion		X (Male)		
Failure to prioritize project	X (Male)			X (Male)
Inconsistency of contribution	X (Female)	X (Male)		
Inconsistency with an engineering identity		X (Male)		
Lack of communication		X (Male)		
Lack of competence, experience, or skills				
Lack of initiative				X (Female)
Procrastination		X (Male)		
Restricting others' work				
Unreliability			X (Male)	

*(Gender) rates others more severely/frequently **(Gender) is rated more severely/frequently

V. CONCLUSION AND SYNTHESIS

When we are asked to judge one another's behavior, that estimation is impacted by more than just an "objective" observance of behavior itself. Factors related to the identity of both the actor and target play a role in shaping how we view one another. In this paper, gender is explored as one factor that may influence how people perceive others. Gender is one factor that can impact the frequency and severity with which we perceive one another, but there may be many others. Evaluations of behavior is not an objective process. Because gender is shown to impact perceptions of behaviors, the respective genders of existing teams matter. For example, the one female on a team with mostly males may be disproportionately severely penalized in peer evaluations for failing to show enough initiative. Though this work focuses on undergraduate engineering classrooms, its findings extend to the workforce and wherever long-term teams are formed. Industries that disproportionately hire men, for example, will likely see more severe evaluations for faults like failing to communicate or contributing inconsistently. These evaluations can have huge implications on an employee's future (or a student's grade) and are to some extent subject to specific aspects of the actor and target's identities. Thus, those with influence should consider forming teams with this principle in mind and seek to balance gender representation within teams and, in a larger sense, within their organizations.

Some behaviors are influenced more by the actor's gender and some by the target's gender. This implies that some behaviors are more subject to the actor's interpretation of the behaviors while other behaviors can be observed objectively fairly easily. Behaviors that can be observed straightforwardly (in terms of either frequency or severity) include lack of initiative and unreliability. Behaviors subject to the actor's interpretation of behaviors include failure to advance project toward completion, inconsistency of contribution, inconsistency with an engineering identity, lack of communication, and procrastination. Failure to prioritize the project is one behavior that had elements of being influenced by both the gender of the actor and target. In particular, male actors rated others as having this behavior more frequently than females did and males who were judged to exhibit this behavior were penalized more severely than the females judged to exhibit the behavior. Thus, experience of this behavior may be co-constructed by the relationship between the pair. One interpretation of this may be that failing to be physically present in group meetings is easily excused when the relationship with that individual is positive while the behavior is judged severely when the relationship between actor and target is poor.

References

[1] L. Springer, M. E. Stanne and S. S. Donovan, "Effects of Small-Group Learning on Undergraduates in Science, Mathematics, Engineering, and Technology: A Meta-Analysis," *Review of Educational Research,* vol. 69, *(1),* pp. 21-51, 1999.

[2] R. Martin *et al*, "Engineering graduates' perceptions of how well they were prepared for work in industry," *European Journal of Engineering Education,* vol. 30, *(2),* pp. 167-180, 2005.

[3] C. Mcdowell *et al*, "The impact of pair programming on student performance, perception and persistence," in *25th International Conference on Software Engineering, 2003. Proceedings.* 2003.

[4] S. E. Cross and N. V. Vick, "The Interdependent Self-Construal and Social Support: The Case of Persistence in Engineering," *Person. Soc. Psychol Bull.,* vol. 27, *(7),* pp. 820-832, 2001.

[5] R. Suresh, "The Relationship between Barrier Courses and Persistence in Engineering," *Journal of College Student Retention: Research, Theory & Practice,* vol. 8, *(2),* pp. 215-239, 2006.

[6] H. Tajfel and J. C. Turner, *The Social Identity Theory of Intergroup Behavior.* 2004.

[7] G. L. Stewart, I. S. Fulmer and M. R. Barrick, "An Exploration of Member Roles as a Multilevel Linking Mechanism for Individual Traits and Team Outcomes," *Person. Psychol.*, vol. 58, *(2)*, pp. 343-365, 2005.

[8] S. Sonnentag and J. Volmer, "Individual-Level Predictors of Task-Related Teamwork Processes: The Role of Expertise and Self-Efficacy in Team Meetings," *Group & Organization Management,* vol. 34, *(1),* pp. 37-66, 2009.

[9] A. Zhang, "Peer Assessment of Soft Skills and Hard Skills," *Journal of Information Technology Education,* vol. 11, *(1),* 2012.

[10] L. E. Gueldenzoph and G. L. May, "Collaborative Peer Evaluation: Best Practices for Group Member Assessments," *Business Communication Quarterly,* vol. 65, *(1),* pp. 9-20, 2002.

[11] S. Fallows and B. Chandramohan, "Multiple Approaches to Assessment: Reflections on use of tutor, peer and self-assessment," *Teaching in Higher Education,* vol. 6, *(2),* pp. 229-246, 2001.

[12] M. Borrego *et al*, "Team Effectiveness Theory from Industrial and Organizational Psychology Applied to Engineering Student Project Teams: A Research Review," *J Eng Educ,* vol. 102, *(4),* pp. 472-512, 2013.

[13] A. Erez, J. A. Lepine and H. Elms, "Effects of Rotated Leadership and Peer Evaluation on the Functioning and Effectiveness of Self-Managed Teams: A Quasi-Experiment," *Person. Psychol.,* vol. 55, *(4),* pp. 929-948, 2002.

[14] D. F. Baker, "Peer Assessment in Small Groups: A Comparison of Methods," *Journal of Management Education,* vol. 32, *(2),* pp. 183-209, 2008.

[15] E. Miller, L. Hirshfield and D. Chachra, "Classifying dissatisfaction: Student perspectives on teammate performance," in *American Society for Engineering Education,* New Orleans, LA, 2016, .

[16] M. C. Thomas and K. W. Phillips, "When What You Know Is Not Enough: Expertise and Gender Dynamics in Task Groups," *Person. Soc. Psychol Bull.,* vol. 30, *(12),* pp. 1585-1598, 2004.

[17] A. Joshi, "By Whom and When Is Women’s Expertise Recognized? The Interactive Effects of Gender and Education in Science and Engineering Teams," *Adm. Sci. Q.,* vol. 59, *(2),* pp. 202-239, 2014.

[18] L. U. Eric and G. L. Cohen, "Constructed Criteria: Redefining Merit to Justify Discrimination," *Psychol. Sci.,* vol. 16, *(6),* pp. 474-480, 2005.

[19] D. V. Van, J. S. Bunderson and A. Oosterhof, "Expertness diversity and interpersonal helping in teams: Why those who need the most help end up getting the least," *Academy of Management Journal,* vol. 49, *(5),* pp. 877-893, 2006.

[20] M. W. Ohland *et al*, "The Comprehensive Assessment of Team Member Effectiveness: Development of a Behaviorally Anchored Rating Scale for Self- and Peer Evaluation," *Amle,* vol. 11, *(4),* pp. 609-630, 2012.

[21] E. Mather. *Facts and Stats*.