Investigating Gender Bias in Peer Evaluation within First-Year Engineering Teams

Dr. Junqiu Wang, University of North Carolina at Charlotte

Dr Wang currently serves as Teaching Associate Professor at Office of Student Development and Success, William States Lee College of Engineering, University of North Carolina Charlotte. He received his Ph.D. in Engineering Education from Purdue University, M.S. in Materials Engineering from Tsinghua University, China and B.S. in Materials Engineering from University of Sci. & Tech. Beijing, China

Mr. Fazel Ranjbar, University of Cincinnati

Fazel Ranjbar is a doctoral candidate in Engineering and Computing Education at the University of Cincinnati and holds a Graduate Certificate in Data Analytics. He combines a background in industrial engineering and over a decade of management experience with advanced statistical methods to evaluate educational programs. His academic research investigates team dynamics and teamwork in undergraduate students. To support this work, he builds automated scoring platforms and applies natural language processing to validate educational assessments. His research aims to inform institutional decisions and improve instructional interventions in higher education.

Investigating Gender Bias in Peer Evaluation within First-Year Engineering Teams

Abstract

This empirical full paper investigates the gender bias in first-year engineering student project team. Teamwork is a critical professional skill that has received increasing attention in engineering education, especially in first-year engineering curricula. Peer evaluation has become an essential tool for assessing teamwork skills within the engineering education community. However, growing evidence shows that students' evaluations of their peers may be influenced by gender and the team's gender composition, raising questions about fairness and bias in using peer evaluation for assessment purposes. This study investigates the impact of gender bias in peer evaluation and perceived teamwork behaviors in first-year engineering teams. Quantitative data were collected from 2,963 students across two cohorts (Fall 2023 and Fall 2024) using Likert-scale peer evaluation scores and fixed-sum grade allocations. To explore possible factors contributing to bias, teams were categorized into three compositions: (1) Male-Dominated (<50% female), (2) Gender-Balanced (=50% female), and (3) Female-Dominated (>50% female).

Data analysis reveals a "Minority Premium" wherein female students in male-dominated teams received significantly higher peer-assigned grades compared to their male counterparts, despite exhibiting statistically equivalent behavioral ratings. Conversely, this grading advantage disappears in gender-balanced and female-dominated teams. Findings from this study provide insights into how gender composition affects peer evaluation outcomes and inform strategies for equitable team formation and bias reduction in team peer evaluation.

Introduction and literature review

The cultivation of teamwork skills has become a cornerstone of modern engineering education, driven largely by accreditation requirements (ABET Outcome 5) and persistent industry demand for engineers with collaborative competence [8]. To assess these competencies within large first-year cohorts, engineering programs heavily rely on peer evaluation instruments. Tools such as CATME and others have become standard practice, enabling instructors to gather individual-level data on student contributions that would otherwise remain invisible in group deliverables [6, 7]. While the efficiency of these instruments are well-documented, their validity relies on the assumption that students can objectively evaluate their peers' contributions. However, emerging scholarship suggests that social identify factors, particularly gender, may significantly skew these perceptions, thereby challenging the equity and fairness of peer evaluation as an assessment instrument [15]. Previous studies have yielded mixed results regarding gender bias, with some suggesting female students are penalized in male-dominated environments, while others find no significant difference [16].

Feminine-coded gender bias in engineering education consistently highlights the influence of role congruity theory, which posits a fundamental dissonance between the stereotypes of a "good engineer", typically masculine-coded as technical soundness; and "good team citizenship", often feminine-coded as cooperative and organizational role [1]. Consequently, female students in engineering teams frequently report being relegated to managerial or secretarial roles, such as writing reports, scheduling meetings or formatting slides, while their male counterparts dominate technical tasks [2]. This division of labor not only limits technical self-efficacy but also distorts peer evaluation outcomes.

The demographic composition of the team acts as a significant moderator for these biases [4, 5]. In the context of engineering education, women in solo status with male-majority teams often report significantly higher levels of isolation and stereotyping compared to their peers in gender-balanced or female-majority configuration [3]. While some studies suggest that achieving gender parity (50% female) can mitigate these bias, conflicting evidence indicates that gender dynamics may shift rather disappear in balanced setting [2].

This research seeks to empirically exam the relationship between gender, team composition and peer evaluation outcomes in a fundamental engineering design course from a large public university. Specifically, this study addresses the following research questions:

1) Are there any significant gender differences in peer evaluation ratings?
2) Whether the compositions of gender of a team shifts the peer evaluation rations between male and female.

Theoretical Framework

This study grounds its investigation of gender bias and team gender composition within a multidimensional framework of team effectiveness. To rigorously assess teamwork effectiveness in first-year engineering students, we utilize a measurement instrument comprised of four established constructs: Interdependency, Goal setting, team Potency and Trust. These constructs were selected because they represent distinct but interrelated antecedents to successful team collaboration. Interdependency is defined as degree to which team members must rely on one another for completing a team assignment [9, 10, 11]. Wageman identified that high task interdependency is a critical driver to team effectiveness as it necessitates structural collaboration rather than isolated individual effort [12]. Goal setting is recognized as a fundamental mechanics for team performance when clearly defined. More challenging goal setting often leads to better communication and reduce process losses and leads to higher performance outcomes [13]. Additionally, clear goals foster better communication and team coordination [14]. The third dimension is potency, which refers to a team's shared belief in its capability to succeed regardless, serving as a measure of collective efficacy and a robust predictor of team effectiveness [15]. Finally, trust is defined as psychological safety, unlike potency, trust represents a belief in safety which Costa argues is essential for reducing monitoring behaviors and encouraging the risk-taking necessary for innovative engineering design [16].

Table 1 The Measurement instrument

Construct	Items	
Interdependency	IN1	Is collaborating well with the team on all in-class and out of the class assignments
	IN2	Is contributing to the team's success by having a clearly assigned roles
	IN3	Is a reliable team member
Goal Setting	GS1	Is often helping the team develop effective strategies
	GS2	Is articulating individual goals that can be achieved with the help of the team.
	GS3	Is actively helping the team establish goals
Potency	PT1	Is helping the team to build a shared confidence in its ability to successfully work together on course assignments.
	PT2	Is often encouraging each team member to believe in the team's ability to succeed no matter what the task
	PT3	Often making the team feel confident in its ability to resolve disagreements.
Trust	TS1	Is always seen as one who is willing to share important information with teammates.
	TS2	Always appears to be willing to have teammates critically review work performed
	TS3	Is rarely seen checking the work of teammates because of confidence in their approach to solving problems.
Grades	Assign grades to each member regarding your contribution. The sum of the grades are 75 points for a team of 3, 100 for a team of 4 and 125 for a team of 5	

To ensure psychometric soundness, the peer evaluation instrument utilizes three measurement items for each of the four dimensions. Additionally, the instrument includes a quantitative Grades component to triangulate the four-dimension peer evaluation instruments. Using a constant sum method, students distribute a grade pool among their teammates based on contributions (e.g., a pool of 75 points for a team of 3, or 100 points for a team of 4). In this system, a score of 25 represents an equal or "average" contribution. Deviations from this baseline serve as a proxy for perceived inequality in effort or performance. A detailed description of the instrument is provided in Table 1.

Data Collection

This study was conducted at a large public University in the Midwest within a fundamental first-year engineering design course. Participants were undergraduate engineering students collaborating in assigned teams on a semester-long design project. The data collection spanned two academic cohorts: Fall 2023 and Fall 2024.

Peer Evaluation Instrument surveys were administrated electronically via Qualtrics Survey platform. Students were required to evaluate every member of their team, including themselves, across two distinct assessment dimensions:

- Behavioral Assessment: Students rated their peers on a validated 12-item instrument designed to measure key team constructs (interdependency, Goal Setting, Potency, and Trust).
- Contribution grade: Utilizing a fixed sum scoring method, students distributed a set pool of “grade points” among their teammates. This distribution was intended to reflect the perceived percentage of each member’s contribution to the team’s collective output.

A screenshot of the peer evaluation interface is presented in Figure 1.

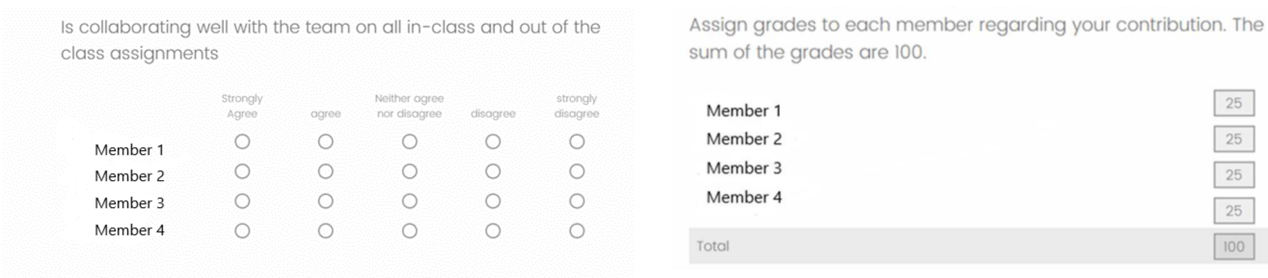

Figure 1 Screenshot of Peer Evaluation Interface that student used for evaluating their peers

Data Analysis

Data was collected from two cohorts of a first-engineering course, fall 2023 and fall 2024. The Fall 2023 cohort includes 1534 students (80% Male, 20.0% Female). The Fall 2024 dataset included 1430 students; however, one record was excluded due to missing gender data, resulting a final analytical sample of 1429 students (77.9% Male, 22.1% Female). Students were assigned into teams of 3-5 members. To analyze the impact of gender diversity, teams were categorized into three compositions based on the proportion of female members: Male-Dominated (<50%), Balanced (= 50% Female), and Female-Dominated (>50% Female). As detailed in Table 2, the distribution of team type shifted between cohorts. In Fall 2023, the majority of teams were Male Dominated (66.7%), with a significant portion being Balanced (24.7%). In Fall 2024, while Male-Dominated teams remained prevalent (75.4%), there was a notable polarization: the proportion of Balanced teams dropped to 7.7%, while Female-Dominated teams nearly doubled to 16.9%.

Table 2 Gender composition profile for Fall 2023 and Fall 2024

Team Composition	Fall 2023	Fall 2024
< **50% Female** (Majority Male)	262 teams (66.7%)	276 teams (75.4%)
= **50% Female** (Balanced)	97 teams (24.7%)	28 teams (7.7%)
> **50% Female** (Majority Female)	34 teams (8.6%)	62 teams (16.9%)

General Gender Effect

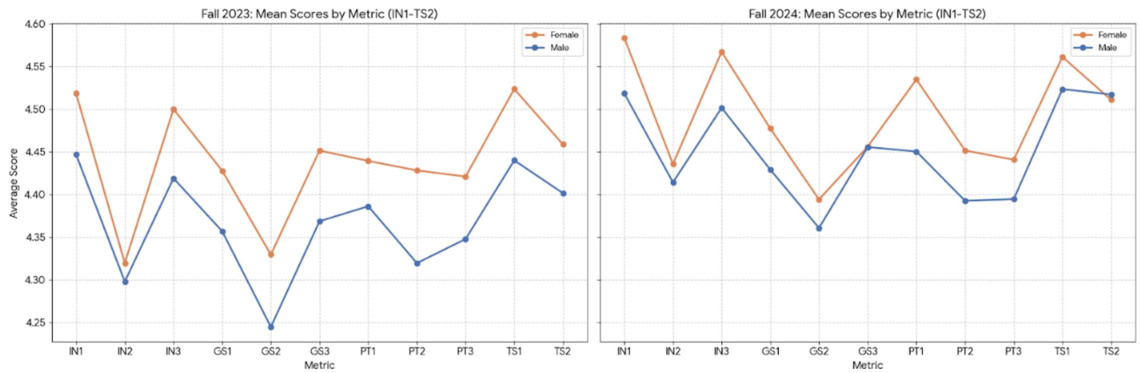

Figure 2 Gender comparison of average ratings for the 11-item peer evaluation instrument (excluding TS3), Fall 2023 and Fall 2024

Figure 2 illustrates the comparison of mean peer evaluation ratings by gender across the 11 validated instrument items. The third Trust item (TS3) was excluded from the analysis due to consistent misinterpretation by first-year students.

The results demonstrate a distinct trend wherein female students received consistently higher ratings than their male counterparts, although the magnitude of these differences is small. Independent T-tests confirmed that these differences were statistically significant for specific items:

- Fall 2023, female students scored significantly higher on PT2 and TS1 ($p < 0.05$).
- Fall 2024 cohort, significance was observed only for PT1 ($p = 0.025$).

The finding indicates that female students are perceived as key drivers of team potency. Regarding team perceived grades or percentage of contribution to teamwork, while female students received slightly higher average (25.48 vs 24.88), the difference did not reach statistical significance in either cohort. The data does not support there is a significant difference between male and female students overall.

Table 3 Item by item comparison of peer evaluation instrument between male and female students, Fall 2023

Item	Male	Female	Diff (F-M)	p-value	Significance
IN1	4.45	4.52	0.07	0.108	False
IN2	4.30	4.32	0.02	0.625	False
IN3	4.42	4.50	0.08	0.082	False
GS1	4.36	4.43	0.07	0.108	False
GS2	4.24	4.33	0.08	0.053	False
GS3	4.37	4.45	0.08	0.056	False
PT1	4.39	4.44	0.05	0.230	False
PT2	4.32	4.43	0.11	0.011	True
PT3	4.35	4.42	0.07	0.087	False
TS1	4.44	4.52	0.08	0.048	True
TS2	4.40	4.46	0.06	0.149	False
Grades	24.88	25.48	0.60	0.107	False

Table 4 Item by item comparison of peer evaluation instrument between male and female students, Fall 2024

Item	**Male**	**Female**	**Diff (F-M)**	**p-value**	**Significance**
IN1	4.52	4.58	0.07	0.101	False
IN2	4.41	4.44	0.02	0.593	False
IN3	4.50	4.57	0.07	0.116	False
GS1	4.43	4.48	0.05	0.210	False
GS2	4.36	4.39	0.03	0.390	False
GS3	4.46	4.46	0.00	0.990	False
PT1	**4.45**	**4.53**	**0.08**	**0.025**	**True**
PT2	4.39	4.45	0.06	0.116	False
PT3	4.39	4.44	0.05	0.224	False
TS1	4.52	4.56	0.04	0.312	False
TS2	4.52	4.51	-0.01	0.857	False
Grades	24.91	24.76	-0.15	0.712	False

The first research question investigated whether female students receive different peer evaluation ratings than their male counterparts. Across both cohorts of students, the data reveals a pattern of statistical parity rather than a broad performance gap. Even though female students receive relative higher peer evaluation scores, only one or two elements out of the 11 peer evaluation instruments show statistical significance. This find is critical as it refutes the historical deficit model, which posits that women in engineering struggle because they lack specific competencies, confidence comparing to male [19, 20]. Interventions based on this model often focus on fixing the women to fit existing engineering culture [20, 21]. However, our data demonstrates that female students are performing on par with their male counterparts, if not better. This statistical equivalence suggests that women are fully integrated contributors, neither lagging nor requiring remedial behavioral interventions.

Gender Composition Effect

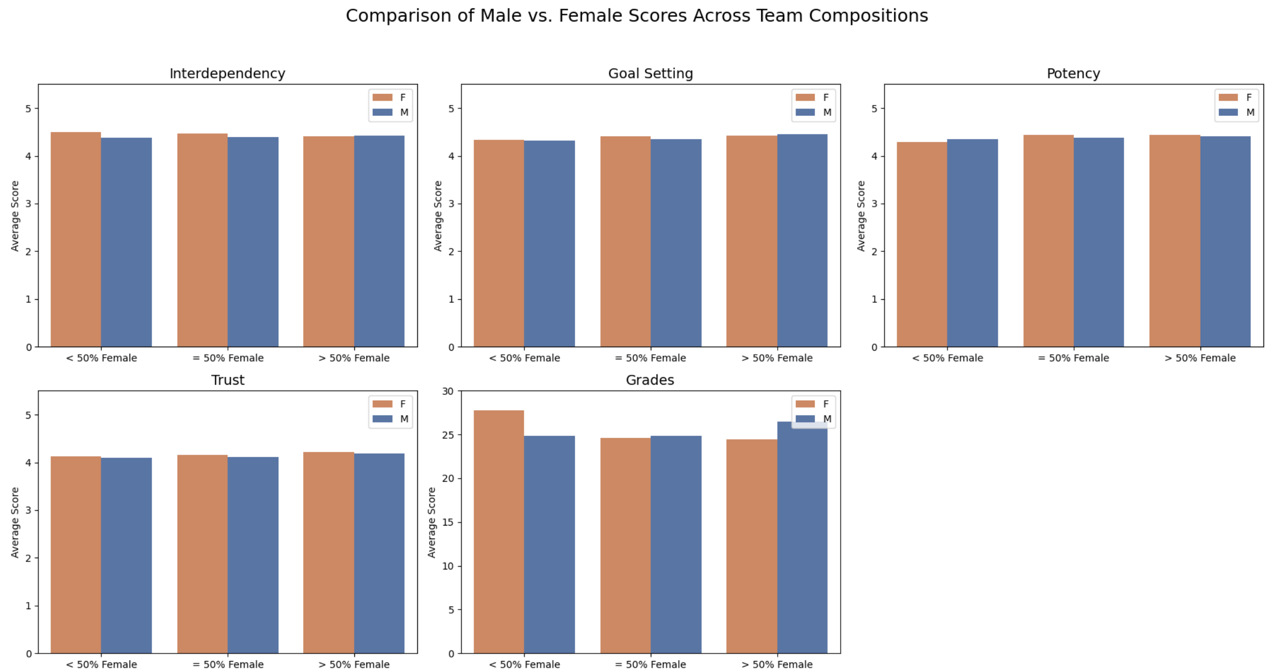

Figure 3 Gender Compositions versus Peer Evaluation and Grades (Fall 2023)

Figure 3 presents the mean peer evaluation scores and peer-assigned average grades for the fall 2023 cohort, segmented by team gender composition: Male-dominated (<50% Female), Gender balanced (=50% female) and Female-dominated (>50% female). As shown in the figure, the behavior anchored constructs: interdependency, goal setting, potency and trust display quite strong stability across team composition. The visual gaps between male and female ratings for these behavioral metrics are minimal across all the three team compositions, numerical data as shown in Table 5 also support this conclusion. The figure also highlights a distinct interaction effect regarding grades. In male-dominated teams(<50% female), female students received a notable grade premium compared to their male counterparts (27.73 vs 24.86), albeit non-significant due to small sample size. Interestingly, this trend reverses in Female-dominated teams (24.48 VS 26.46). Detailed data available in Table 5.

The data analysis results suggest that while peers evaluate behaviors similarly regardless of team makeup, the assignment of grades tends to favor the gender minority: women are rewarded more in male dominated group and men are rewarded more in female dominated group.

Table 5 Impact of Team Gender Composition on Peer Evaluation (Statistics, Fall 2023)

Team Composition	Metric	Male Mean	Female Mean	Difference (F-M)	P-Value	Significant?
< 50% Female (Male-Dominated)	Grades	24.86	27.73	2.86	0.102	False
	Interdependency	4.39	4.50	0.12	0.255	False
	Goal Setting	4.31	4.33	0.02	0.870	False
	Potency	4.34	4.29	-0.06	0.664	False
	Trust	4.10	4.13	0.03	0.810	False
= 50% Female (Gender-Balanced)	Grades	24.84	24.63	-0.21	0.745	False
	Interdependency	4.39	4.46	0.07	0.356	False
	Goal Setting	4.35	4.40	0.05	0.457	False
	Potency	4.37	4.43	0.06	0.371	False
	Trust	4.12	4.16	0.05	0.481	False
> 50% Female (Female-Dominated)	Grades	26.46	24.48	-1.99	0.180	False
	Interdependency	4.42	4.40	-0.02	0.897	False
	Goal Setting	4.45	4.42	-0.03	0.783	False
	Potency	4.41	4.44	0.03	0.773	False
	Trust	4.19	4.21	0.02	0.843	False

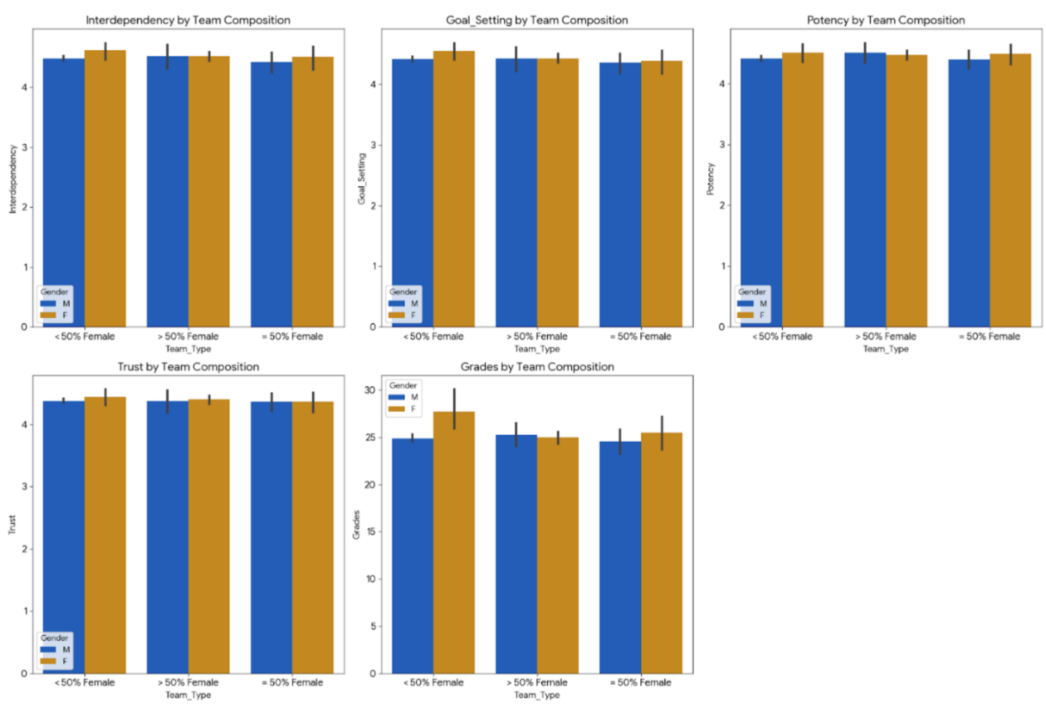

Figure 4 Team gender Compositions and Peer Evaluation and Grades (Fall 2024)

Table 6 Impact of Team Gender Composition on Peer Evaluation (Statistics, Fall 2023)

Team	Metric	Male	Female	Difference	P-Value	Significant?
< 50% Female (Male-Dominated)	Grades	24.89	27.72	2.83	0.012	True
	Interdependency	4.48	4.61	0.13	0.092	False
	Goal Setting	4.42	4.55	0.14	0.07	False
	Potency	4.41	4.5	0.09	0.222	False
	Trust	4.38	4.44	0.06	0.353	False
= 50% Female (Gender-Balanced Teams)	Grades	24.53	25.47	0.94	0.406	False
	Interdependency	4.42	4.51	0.08	0.521	False
	Goal Setting	4.36	4.39	0.04	0.774	False
	Potency	4.4	4.48	0.09	0.441	False
	Trust	4.37	4.37	0	0.996	False
> 50% Female (Female-Dominated)	Grades	25.23	24.97	-0.26	0.716	False
	Interdependency	4.52	4.52	0	0.978	False
	Goal Setting	4.43	4.43	0.00	0.97	False
	Potency	4.5	4.47	-0.04	0.685	False
	Trust	4.37	4.4	0.03	0.758	False

Figure 4 and Table 6 present the data analysis results for Fall 2024 cohort. The integration of Fall 2024 data with the Fall 2023 cohort reveals a striking and statistically robust pattern: the gender composition of a team significantly moderates how female students are evaluated, specifically when they are in the numerical minority.

1. The Minority Premium in Male-Dominated Teams

The most critical finding across both years is the performance of female students in Male-Dominated teams (<50% female).

- Consistency of Effect: In Fall 2024, female students in these teams received significantly higher peer-assigned grades than their male teammates (27.72 VS 24.89), with a difference of 2.83 points (p value of 0.012). This almost perfectly mirrors the Fall 2023 data, where the grade difference was 2.86 points.
- Statistical Confirmation: While the Fall 2023 difference failed to reach significance (p=0.102) due to a small sample size of women in this subgroup (N=18), the Fall 2024 dataset confirms that this is a statistically real phenomenon (N = 48). This suggests that female students are not penalized in a male-dominated team; rather, they are consistently recognized as high contributors in male-majority environments.

2. Contextual Fading of the Female Advantage

The data further indicates that this grading advantage is context dependent. As the proportion of women in the team rises, the premium disappears:

- Gender-Balanced Teams: In Fall 2024, the grade difference shrank to a non-significant 0.94 points (p=0.406), consistent with the negligible -0.21 difference observed in Fall 2023.

- Female-Dominated Teams: When women comprised the majority (>50%), the advantage vanished entirely. In Fall 2024, male and female grades were effectively equal (Difference = -0.26, p=0.716). This echoes the Fall 2023 trend where the gap reversed, favoring men (Difference = -1.99), though neither reversal was statistically significant.

3. Behavioral Stability vs. Grading Volatility

A distinct contrast exists between Grades (perceived contribution to the teamwork) and Behavioral Ratings (Interdependency, Goal Setting, Potency, Trust).

- While grades fluctuated significantly based on team makeup, behavioral scores remained remarkably stable. In Fall 2024, female students in male-dominated teams scored higher on Interdependency (0.13) and Goal Setting (0.14), but these gaps did not widen dramatically compared to other team types.
- This divergence suggests that while peers *perceive* female behaviors similarly across contexts (consistently positive), they feel compelled to *reward* those behaviors more heavily (via grades) when the female student is a distinct minority.

Discussion and Conclusion

These findings have practical implications for how engineering educators structure teams and assess gender dynamics. Educators are often advised to avoid placing women as "solos" in teams to prevent isolation. However, our data suggests that dispersing female students into male-majority teams does not inherently disadvantage them in terms of assessment; in fact, these environments currently serve as the primary driver of recognition for female students' contributions. While social belonging remains a concern, performance devaluation does not appear to be the primary risk. The disconnect between Behavior (no significant differences) and Grades (biased towards female in minority compositions) suggests that "fixed-sum" grading schemes are susceptible to demographic bias. Moving toward assessment models based on anchored behavioral ratings—where men and women scored equally—may yield fairer outcomes than subjective grade allocations.

Limitation of the study

While this study benefits from a multi-year dataset (Fall 2023 and Fall 2024), the analysis is primarily quantitative. Although open-ended qualitative feedback was collected from students regarding their peers' positive and negative behaviors during fall 2024, a comprehensive thematic analysis of these narratives was outside the scope of the current quantitative investigation. The unanalyzed qualitative dataset presents a rich opportunity for future work. Our next phase of research will utilize Natural Language Processing (NLP) to triangulate the quantitative findings presented here.

Acknowledgment

This work was supported by KEEN. Google Gemini is used for language and grammar editing. The use of data for the research is approved by University of North Carolina at Charlotte, IRB-26-0498.

Reference

[1] A. H. Eagly and S. J. Karau, “Role congruity theory of prejudice toward female leaders,” *Psychological Review*, vol. 109, no. 3, pp. 573–598, 2002.

[2] L. A. Meadows and D. Sekaquaptewa, “The influence of gender stereotypes on role adoption in student teams,” in *Proc. ASEE Annu. Conf. Exposition*, Atlanta, GA, USA, 2013.

[3] N. Dasgupta, M. Scircle, and M. Hunsinger, "Female peers in small work groups enhance women's motivation, verbal participation, and career aspirations in engineering," *Proceedings of the National Academy of Sciences*, vol. 112, no. 16, pp. 4988-4993, 2015.

[4] W. C. Lee and H. M. Matusovich, “A mixed-methods exploration of equity and inclusion in engineering teamwork,” *Journal of Engineering Education*, vol. 105, no. 4, pp. 566–590, 2016.

[5] D. B. Kaufman, R. M. Felder, and H. Fuller, “Peer ratings in collaborative work groups: Effects of gender and ethnicity,” *Journal of Engineering Education*, vol. 89, no. 1, pp. 133–140, 2000.

[6] R. Layton, M. Ohland, and H. Pomeranz, “Software for student team formation and peer evaluation: CATME incorporates Team-Maker,” in *Proc. ASEE Annu. Conf. Exposition*, Honolulu, HI, USA, 2007.

[7] M. Loughry, M. Ohland, and D. D. Moore, “Development of a theory-based assessment of team member effectiveness,” *Educational and Psychological Measurement*, vol. 67, no. 3, pp. 505–524, 2007.

[8] R. M. Felder, D. R. Woods, J. E. Stice, and A. Rugarcia, “The future of engineering education II: Teaching methods that work,” *Chemical Engineering Education*, vol. 34, no. 1, pp. 26–39, 2000.

[9] M. A. Campion, G. J. Medsker, and A. C. Higgs, “Relations between work group characteristics and effectiveness: Implications for designing effective work groups,” *Personnel Psychology*, vol. 46, no. 4, pp. 823–847, 1993.

[10] J. R. Hackman, “The design of work teams,” in *Handbook of Organizational Behavior*, J. W. Lorsch, Ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 1987, pp. 315–342.

[11] A. Somech, H. S. Desivilya, and H. Lidogoster, “Team conflict management and team effectiveness: The effects of task interdependence and team identification,” *Journal of Organizational Behavior*, vol. 30, no. 3, pp. 359–378, 2009.

[12] R. Wageman, “Interdependence and group effectiveness,” *Administrative Science Quarterly*, vol. 40, no. 1, pp. 145–180, 1995.

[13] M. Hoegl and K. P. Parboteeah, “Goal setting and team performance in innovative projects: On the moderating role of teamwork quality,” *Small Group Research*, vol. 34, no. 1, pp. 3–19, 2003.

[14] A. Kleingeld, H. van Mierlo, and L. Arends, “The effect of goal setting on group performance: A meta-analysis,” *Journal of Applied Psychology*, vol. 96, no. 6, pp. 1289–1304, 2011.

[15] J. Hu and R. C. Liden, “Antecedents of team potency and team effectiveness: An examination of goal and process clarity and servant leadership,” *Journal of Applied Psychology*, vol. 96, no. 4, pp. 851–862, 2011.

[16] A. C. Costa, “Work team trust and effectiveness,” *Personnel Review*, vol. 32, no. 5, pp. 605–622, 2003.

[17] K. A. Jehn, “A multimethod examination of the benefits and detriments of intragroup conflict,” *Administrative Science Quarterly*, vol. 40, no. 2, pp. 256–282, 1995.

[18] L. A. DeChurch and J. R. Mesmer-Magnus, “The cognitive underpinnings of effective teamwork: A meta-analysis,” *Journal of Applied Psychology*, vol. 95, no. 1, pp. 32–53, 2010.

[19] K. Beddoes, “Engineering education discourses on underrepresentation: Why problematization matters,” *Int. J. Eng. Educ.*, vol. 27, no. 5, pp. 1117–1129, 2011.

[20] S. L. Sheffield, “Women in engineering: To the rescue,” in *Women and Science: Social Impact and Interaction*, S. L. Sheffield, Ed. New Brunswick, NJ, USA: Rutgers Univ. Press, 2004, pp. 125–142.

[21] C, Zhou, S. Choi, B Beigpourian, S. Wei, D. M. Ferguson, M. Ohland, “The Difference between Teams with No Female Students and Teams with Female Students for Peer Evaluation Behavior in Engineering Education” Proceedings of the International Conference on Industrial Engineering and Operations Management Toronto, Canada, October 23-25, pp 609 -619, 2019