Bias in First-Year Engineering Student Peer Evaluations

Lea Wittie, Bucknell University

Lea Wittie is an Associate Professor in the department of Computer Science in the Engineering College at Bucknell University. She has spent the past 4 years coordinating the first year Engineering student Introduction to Engineering and over a decade participating in the program before that.

James Bennett, Cornell University

James Bennett is a biomedical engineer specializing in medical device design and development. He has earned a Bachelor of Science Degree in Biomedical Engineering from Bucknell University and is currently pursuing a Master's of Engineering in Biomedical Engineering at Cornell University.

Carly Merrill, Bucknell University

Carly Merrill is currently working in the healthcare industry where she is pursing a career in strategic product development. She has recently earned a Bachelor of Science Degree in Biomedical Engineering from Bucknell University.

Dr. Jove Graham, Geisinger

Jove Graham, PhD is an Associate Professor in the Center for Pharmacy Innovation and Outcomes at Geisinger, a nonprofit integrated health system in Pennsylvania.

Troy Schwab, Bucknell University

Troy Schwab is a computer scientist currently working as a federal consultant, specifically concerning data engineering. He received undergraduate degrees in computer science and contemporary music composition from Bucknell University.

Bias in First Year Engineering Student Peer Evaluations

Lea Wittie
Department of Computer Science
Bucknell University
lwittie@bucknell.edu

James Bennett
Cornell University
jameshbennettjr@gmail.com

Carly Merrill
Bucknell University
cpm013@bucknell.edu

Jove Graham
Center for Pharmacy Innovation & Outcomes
Geisinger
jhgraham1@geisinger.edu

Troy Schwab
Bucknell University
trschwab7@gmail.com

Abstract

This Complete Research paper describes a study on race, gender, and self-bias in first year engineering student's team peer evaluations. Our institution runs a first year introduction to engineering course with approximately 200 students that uses team projects over the span of the semester. Each project has 2-5 students per team and incorporates peer and self evaluations into each student's individual project grades. The researchers began this study to observe how racial, gender, and self-bias impact these peer evaluations. Peer evaluations are often employed in instances of group work, particularly in the undergraduate setting. These peer evaluations can present important information regarding team dynamic and distribution of workload. However, this method of assessment is also susceptible to both explicit and implicit biases, specifically in regard to race, gender, and self-bias. After identifying possible biases in our peer evaluation procedure, the researchers plan to examine methods to mitigate these biases. For each project, students submitted peer evaluations of themselves and each of their team members. The peer evaluations required the students to split 100 points amongst all team members including themselves with an optional written rationale for scores. The 1725 peer evaluation scores collected by this study were double-key entered into a database. Participants were also asked to self-identify as one of 6 gender options and 8 race options. If participants selected multiple races, they were assigned to the less common one. Two generalized linear regression models (GLMs) were used, one to estimate self-bias within each race/gender group (i.e. whether students give higher scores to themselves), and one to estimate how members in each race/gender group scored members of other groups, excluding the self-scores. Model coefficients significantly different

from zero at the $p<0.05$ level indicated differences between groups and therefore possible evidence of biases. There were 160 participants, all of whom identified as either male or female. Due to small numbers, participants were combined into 3 race categories (White, Asian, or 'Other') for a total of 6 race/gender groups. Results showed that the students were significantly more likely to give themselves a higher score than other students on average, even after accounting for race/gender. This self-bias was consistent in both genders. White males gave scores to Other males that were significantly lower than what they gave to all other groups and significantly lower than what Other males received from White females or Asian males. This suggests a possible negative bias from White males to Other males. Similarly, Asian males gave scores to Other males and Other females that were significantly higher than what they gave to all other groups and significantly higher that what Other males and Other females received from all other groups. This suggests a possible positive bias from Asian males to Other males and Other females.

Introduction

In undergraduate university curricula, specifically engineering, ability to work in teams is a proficiency which is sought to be instilled within all students. The ability to serve as a valuable team member is necessary across many professional fields and for ensuring students develop skills in collaboration within a team, including understanding diverse perspectives and problem solving approaches. In order to corroborate the efforts of each member of a team of students in a teaching setting, a method of peer evaluations is often employed. These peer evaluations can be utilized by instructors to gain a greater understanding of the dynamic of a team, and sometimes influence the grading of students in the class[1]. It is noted across literature that the method of peer evaluations is met with both praise and adversity from both educators and students, as it gives students the opportunity to share opinions on the dedication of fellow team members, but also presents opportunities for unfair assessments[2?,3,4,5]. It is undeniable that peer evaluations are susceptible to biases. Specially, there are both implicit and explicit biases, both of which may result in assessing peers in a partial manner, deviant of the actual team dynamic or work done by each member[6]. Biases present in peer evaluations have been studied in the past, with research indicating gender, race, social style, and self-bias present in different learning environments, or no bias present at all[7,8,9,10,11,12]. This study aims to evaluate the presence or absence of bias existing between team members and concerning race and gender within a first-year level introductory engineering course. Additionally, the study seeks to identify instances of self-bias evident in peer evaluations, in which a person may adjust their self-score based on either positive or negative bias toward themselves. Upon identifying clear occurrences of bias in the peer evaluation procedure, changes could be made in order to mitigate these biases, or at least reduce their effects. The algorithm employed in this study could be utilized again in order to measure the relative success of these adjustments in the future.

Methods

This study follows 160 out of 192 first-year engineering students (and a few arts & science students) in a introduction to engineering course. The students participated in five team projects. After completing each project, each student completed a peer evaluation of all team members,

including themselves. The peer evaluation involves distributing 100 points between themselves and their other team members based on their perceived performance of each member. Team size ranged from two to five people with most teams consisting of four people. Team members were not constant for each of the five projects.

The peer evaluations written by participants were anonymized using identification numbers. Non-participants mentioned in a peer evaluation were anonymized to a general NP and for study purposes. Nothing is known about the givers or recipients of those evaluations. The peer evaluation data was double-key entered into a database.

There were 234 different teams participating and 1725 overall evaluation scores collected in this study.

Each peer evaluation is a table of participants and the score each received (as seen in Table 1). The participant in the top row gave all the scores, therefore the top score is a self-score. A team with 4 members would therefore generate 4 peer evaluation sheets/tables. The scores sum to 100 in each peer evaluation. The expected score for a standard team of 4 is therefore 25 per person. As team sizes ranged from 2 to 4 members, we scaled scores so that each expected score is 25 even for smaller teams.

Table 1: A sample peer evaluation

Participant	Score given
1	35 (self-score)
2	20
3	35
4	10

We collected gender and racial demographics of the students. The gender options (Male, Female, Trans male/ Trans man, Trans female/ Trans woman, and Genderqueer/Gender non-conforming) along with an additional option to write in a gender identity conforms to best practices[13].

The students in this study self reported only the male and female gender identities. As bias in peer evaluations would be based on public perception, an “in the closet” person would likely be publicly perceived as the dominant categories of male and female, thus this issue should not affect our study results.

This study included students who identified as White, Asian, Black/African American, Hispanic/Latino/Spanish, Middle Eastern/North African, Turkman, Other, and combinations of these groups. Due to small sample size, every racial or ethnic group that was not White or Asian was counted as Other. If participants selected multiple races, they were assigned to the less common one.

A summary of the race and gender demographic of the participants is displayed in Table 2.

We used two different generalized linear regression models to identify evidence of bias when self-evaluating and biases among race/gender groups. Model 1 used 12 terms to estimated mean scores as a function of the 6 race/gender groups and the interaction of self-evaluation across these 6 groups, in the formula shown in Eq. 1.

Table 2: Race and gender demographics of study participants

	White	Asian	Other	Total
Male	77 (48%)	17 (11%)	14 (9%)	108 (68%)
Female	39 (24%)	7 (4%)	6 (4%)	52 (33%)
Total	116 (73%)	24 (15%)	20 (13%)	160 (100%)

$$\text{Score} = B_0 + B_1(\text{White Female}) + B_2(\text{Asian Male}) + B_3(\text{Asian Female}) + B_4(\text{Other Male}) + B_5(\text{Other Female}) + B_6(\text{Self}) + B_7(\text{Self*White Female}) + B_8(\text{Self*Asian Male}) + B_9(\text{Self*Asian Female}) + B_{10}(\text{Self*Other Male}) + B_{11}(\text{Self*Other Female})$$

Figure 1: Model 1 formula using White males as the reference group.

In Model 2, self-evaluation scores were removed from the data, and 36 terms were used instead to estimate the 36 mean scores associated with pairwise combinations of each of the 6 race/gender groups evaluating other students in every other group, in the formula shown in Eq. 2.

Table 3: Sample size (N) in Receiver group, excluding self-evaluation scores

Giver Group	White males	White females	Asian males	Asian females	Other males	Other females
White males	270	135	55	21	51	22
White females	142	75	24	19	28	14
Asian males	53	23	17	4	9	3
Asian females	23	17	3	0	5	2
Other males	40	22	7	5	2	2
Other females	22	14	3	2	6	0

The sample sizes for the non-self evaluation scores used in both model 1 and model 2 are seen in Table 3.

Estimates of the corresponding means, standard errors, t-values, and p-values for the two models were generated and examined to look for statistically significant differences in means that could indicate biases.

Results

Our results are given in Tables 4 and 5 as *estimates and 95% confidence intervals* for the mean scores given from one race/gender group to another. P-values smaller than 0.05 can be interpreted as the difference in means between two groups being statistically significant, that is, having a very low probability ($<5\%$) of being zero. The 95% confidence intervals can be interpreted as the range of hypothesized estimates for the value that are supported by the data and would not be rejected by statistical testing.

Table 4 shows the results for model 1. All groups except for Asian females show significant positive self-bias. The mean amount of self-bias seen in Other females was only slightly significant and also very small as it ranged from negative to positive biases. White students and

Score = B_0 + B_1(White Female Receiver) + B_2(Asian Male Receiver) + ...
+ B_7(White Female Giver) + B_8(Asian Male Giver) + ...
+ B_{12}(White Female Giver*White Female Receiver)
+ B_{13}(White Female Giver*Asian Male Receiver) + ...
+ B_{36}(Other Female Giver*Other Female Receiver)

Figure 2: Model 2 formula using White males as the reference group.

Table 4: Model 1: self-scores vs scores given by someone else.

Group	Mean Score Received (not from self) [95% CI]	Mean Score Received (from self) [95% CI]	Mean difference between self-evaluations and non-self evaluations [95% CI]	p-value
White males	24.7 [24.2, 25.1]	27.0 [26.4, 27.6]	2.3 [1.6, 3.1]	<0.0001
White females	24.5 [23.9, 25.1]	26.9 [26.1, 27.7]	2.4 [1.4, 3.4]	<0.0001
Asian males	23.9 [22.9, 24.9]	26.2 [24.9, 27.5]	2.3 [0.6, 3.9]	0.007
Asian females	25.8 [24.4, 27.2]	26.3 [24.3, 28.3]	0.5 [-1.9, 3.0]	0.67
Other males	23.4 [22.4, 24.4]	26.9 [25.3, 28.5]	3.5 [1.6, 5.4]	0.0003
Other females	25.7 [24.2, 27.2]	25.8 [23.7, 27.8]	0.1 [-2.5, 2.7]	0.017

male students in general showed a clear propensity for about a 2 point self-bias which, given the expected score of 25, means they gave themselves 8% higher scores than they gave other students. White students in particular showed a very strong probability of positive self-bias.

Table 5 shows the results for model 2. Figure 3 shows the mean scores in chart form. There were 3 statistically significant results; low scores given by White males to Other males, high scores given by Asian males to Other males, and high scores given by Asian males to Other females. These can be seen in Figure 3 as the low bar in the White males giver group and the two high bars in the Asian males giver group.

Discussion

The results in model 1 showed clear evidence of self-bias. Some amount of self-bias would indicate a healthy self-respect among our students. This self-bias is also, anecdotally, seen in the course evaluations the students fill out each semester where across the board, they all indicate that they worked harder than their classmates. The more prominent issue is that the bias observed here is nonuniform across all groups of race and gender. We also encountered anecdotal evidence from the students of other students who would lower the scores of their teammates in order to inflate their own score. The perception of this biased activity caused as much harm, if not more, than the actual self-bias did. Student concerns have been documented in other studies[14,15] and make students less willing to participate in the peer evaluation process.

In response to the issue of nonuniform self-bias and the perception of bias lowering other people's scores, we changed our peer evaluation process so that the scores given were each out of 100 instead of all summing to 100 and thus were unlinked from the rest of the scores in an evaluation.

Table 5: Model 2: Mean scores received by group [95% CI]

	Receiver group					
Giver group	White males	White females	Asian males	Asian females	Other males	Other females
White males	24.7 [24.1, 25.3]	24.3 [23.6, 25.0]	24.0 [22.2, 25.8]	25.4 [24.5, 26.2]	21.6[1] [20.0, 23.1]	24.3 [23.4, 25.2]
White females	24.4 [23.7, 25.0]	24.9 [24.1, 25.6]	23.6 [22.2, 24.7]	26.3 [24.9, 27.7]	24.0 [21.5, 26.6]	23.8 [22.4, 24.9]
Asian males	25.2 [24.1, 26.3]	24.6 [23.4, 25.6]	24.1 [22.0, 26.2]	25.5 [20.7, 30.3]	30.2[2] [27.1, 33.4]	43.5[3] [38.0, 49.0]
Asian females	25.1 [23.8, 26.4]	24.7 [22.2, 26.3]	23.3 [17.8, 28.7]	[4]	25.9 [21.6, 30.2]	27.4 [20.6, 34.1]
Other males	24.2 [22.6, 25.3]	23.6 [21.3, 25.5]	24.7 [23.7, 25.5]	25.8 [21.5, 30.1]	25.0 [18.3, 31.7]	26.0 [19.3, 32.7]
Other females	25.7 [24.6, 27.0]	24.6 [22.8, 25.9]	22.5 [17.0, 28.0]	26.3 [19.5, 33.0]	23.4 [19.5, 27.3]	[5]

1. The mean score given by White males to Other males was significantly lower than the mean scores given by White males to all other groups (White males $p<0.0001$, White females $p=0.0007$, Asian males $p=0.01$, Asian females $p=0.003$, Other females $p=0.03$), and significantly lower than the mean scores given to Other males by White females ($p=0.03$).
2. The mean score given by Asian males to Other males was significantly higher than the mean scores given by Asian males to White males ($p=0.004$), White females ($p=0.003$), or Asian males ($p=0.002$), and significantly lower than the mean scores given to Other males by White males ($p<0.0001$), White females ($p=0.0009$) or Other females ($p=0.008$).
3. The mean score given by Asian males to Other females was significantly higher than the mean scores given by Asian males to all other groups (all were $p<0.0001$), and significantly higher than the mean scores given to Other females by White males ($p<0.0001$), White females ($p<0.0001$), Asian females ($p=0.0002$) or Other males ($p<0.0001$).
4. Sample size of 0 for non-self evaluation scores given from Asian females to Asian females.
5. Sample size of 0 for non-self evaluation scores given from Other females to Other females.

Behind the scenes, we altered our usage of the scores as to give self scores less weight than scores given to others. Future work will include a follow-up study to see if the unlinking has lowered the effect of self-bias on scores given to other students.

The results in model 2 showed a negative race/gender bias from White males to Other males. As the Other category includes Black, Hispanic, and Middle Eastern males, groups that have typically experienced bias from White individuals[16], this is not surprising. It is more surprising that the results showed a positive bias from Asian males toward Other males and Other females given past studies on inter-racial relationships[17]. Overall, studies[18,19] have shown that students handle peer evaluations better when given guidance on formulating their responses. Although we do provide some guidance, a more structured approach could result in less bias in the future.

We, like many Universities, are continually updating our teaching of diversity related issues. We already cover the basic idea that diversity brings a wider array of ideas which leads to a wider set

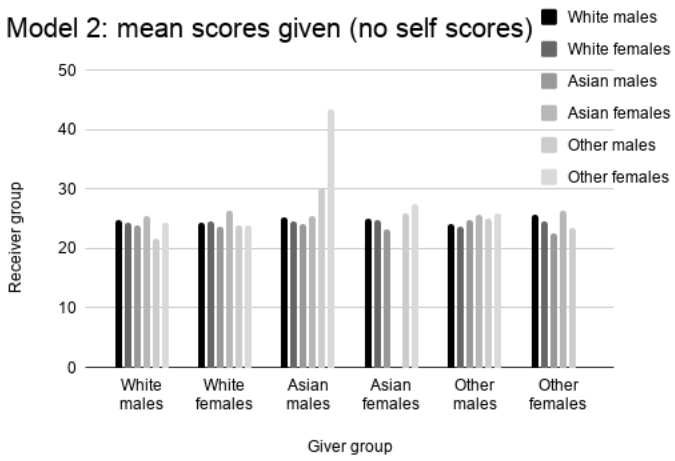

Figure 3: Model 2: mean scores give (no self scores)

of possible solutions. As these are first-year students who are just beginning to know themselves, let alone confront the idea that they may hold implicit biases, we increased the number of activities designed to let the students introspect about themselves and their own traits. We also hope that the general education requirements, themselves continually updating, will bring further growth in this area.

References

[1] K.S. Double, J.A. McGrane, and T.N. Hopfenbeck. The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educ Psychol Rev 32*, 2020.

[2] Martin Fellenz. Toward fairness in assessing student groupwork: A protocol for peer evaluation of individual contributions. *Management Education*, 30(4):570, 2006.

[3] Keith Topping. Self and peer assessment in school and university: Reliability, validity, and utility. *Optimising New Modes of Assessment: In Search of Qualities and Standards. Innovation and Change in Professional Education*, 1:55, 2003.

[4] Robert Thompson. Reliability, validity, and bias in peer evaluations of self-directed interdependent work teams. *ASEE Annual Conference*, 2001.

[5] Peter Ostafichuk and Jim Sibley. Self-bias and gender-bias in student peer evaluation: An expanded study. *Canadian Engineering Education Association (CEEA-ACEG19) Conference*, 2019.

[6] Jacklin Stonewall, Michael Dorneich, and Cassandra Rongerude. A review of bias in peer assessments. *The Collaborative Network for Engineering and Computing Diversity Conference, ASEE*, 2018.

[7] Corrine Moss-Racusin, John Dovidio, Victoria Brescoll, Mark Graham, and Jo Handelsman. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences of the United States of America PNAS*, 2012.

[8] Lisa Gueldenzoph and Gary May. Collaborative peer evaluation: Best practices for group member assessments. *Business Communication Quarterly*, 65(1):9, 2002.

[9] Charlotte Rees. Self assessment scores and gender. *Medical Education*, 37:572, 2003.

[10] Samir Haffar, Fateh Bazerbachi, and M. Hassan Murad. Peer review bias: A critical review. *Mayo Clinic Proceedings*, 94(4):670, 2019.

[11] Douglas Magin. Reciprocity as a source of bias in multiple peer assessment of group work. *Studies in Higher Education*, 26:53–63, 03 2001.

[12] Risna Izati. The influence of friendship bias toward peer assessment in efl classroom. *RETAIN*, 6(2), 2018.

[13] M.V. Lee Badgett, Kellan Baker, Kerith Conron, Gary Gates, Alison Gill, Emily Greytak, and Jody Herman. Best practices for asking questions to identify transgender and other gender minority respondents on population-based surveys. *The GenIUSS Group. The Williams Institute*, 2014.

[14] William Sherrard, Feraidoon Raafat, and Richard Weaver. An empirical study of peer bias in evaluations: Students rating students. *Journal of Education for Business*, 70(1):43–47, 2010.

[15] Heather Verkade and Robert Bryson-Richardson. Student acceptance and application of peer assessment in a final year genetics undergraduate oral presentation. *Journal of Peer Learning*, 6:1, 2013.

[16] Rich Morin. Exploring racial bias among biracial and single-race adults: The IAT. *Pew Research Center, Washington, D.C.*, 2015.

[17] Taking America's pulse: A summary report of the national conference survey on inter-group relations. *National Conference of Christians and Jews, New York, N.Y.*, 1992.

[18] Mariya Omelicheva. Self and peer evaluation in undergraduate education: Are promises worth risking the perils? *Journal of Political Science Education*, 1(1):191, 2005.

[19] Nancy Falchikov and Judy Goldfinch. Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, 70(3):287–322, 2000.