Unconscious Bias in Peer Ratings of International Students' Contributions to First-Year Design Projects?

Dr. Angela R Bielefeldt, University of Colorado, Boulder

Angela Bielefeldt is a professor at the University of Colorado Boulder in the Department of Civil, Environmental, and Architectural Engineering (CEAE) and Director for the Engineering Plus program. She has served as the Associate Chair for Undergraduate Education in the CEAE Department, as well as the ABET assessment coordinator. Professor Bielefeldt was also the faculty director of the Sustainable By Design Residential Academic Program, a living-learning community where students learned about and practice sustainability. Bielefeldt is also a licensed P.E. Professor Bielefeldt's research interests in engineering education include service-learning, sustainable engineering, social responsibility, ethics, and diversity.

Unconscious Bias in Peer Ratings of International Students' Contributions to First-Year Design Projects?

Abstract

Peer ratings are often used to help award individual grades from team projects. It is therefore important to understand the extent to which these peer ratings may be influenced by unconscious or implicit bias. Unconscious bias during peer ratings is perhaps more likely to occur among first-year students who may have previously had limited interactions with diverse groups, in particular international students. To explore this issue, about 600 student peer ratings in the Comprehensive Assessment of Team Member Effectiveness (CATME) system were examined. The students were participating in a team bridge design project in a first-year civil engineering course. Teams worked together for 3 to 5 weeks and submitted a single written report. After the project was completed, students rated themselves and their peers in CATME. Male and female students were rated similarly overall, as were Hispanic and majority students. However, international students from Middle East countries and China had lower adjustment factor ratings than majority peers. The largest differences were in the "contributing" category. Middle Eastern students were equally likely to have designed the 'best bridge' on their teams. This is counter evidence to the lower contributing score awarded to Middle Eastern students by their peers and implies that bias may have been present in the peer ratings. In addition, it was found that male Middle Eastern students rated females lower than males. The results imply that first-year courses that include team design projects should spend time discussing unconscious bias and cultural styles. Faculty should also consider these issues when using peer ratings to assign students individual grades from team projects.

Background

Teamwork is an important skill for engineers, recognized by inclusion among the student learning outcomes in ABET [1]. Teamwork incorporates an array of skills including verbal communication, listening, and respectful collaboration. Being a good teammate also encompasses a variety of attitudes including honesty, open mindedness, tolerance, diligence, reliability, and being considerate [2, pg. 147]. Engineering teams may include individuals from diverse demographic groups (gender, race/ethnicity), as well as an array of characteristics that are not visible. Cultural differences among teammates may be significant as engineering becomes increasingly globalized [2, 3]. The American Society of Civil Engineers included in its Body of Knowledge (CEBOK3) [4] among the professional knowledge and skills that students should possess upon graduation the ability to "Apply concepts and principles of teamwork and leadership, including diversity and inclusion, in the solutions of civil engineering problems." In the rationale, the CEBOK3 notes, "Engineers frequently work in teams…. This requires … being able to foster inclusion of diverse perspectives, cultural backgrounds, knowledge, and experience…" [4, p. 47].

Design projects in engineering courses are commonly executed via teams of students [5]. Peer ratings are commonly associated with team projects, which may help facilitate individual accountability thereby contributing to functional teams. The Comprehensive Assessment of Team Member Effectiveness (CATME) was developed as a behaviorally anchored peer

assessment tool for engineering teams [6,7]. Students rate themselves and their teammates on five aspects of teamwork: contributing to the team's work, interacting with teammates, keeping the team on track, expecting quality, and having relevant knowledge, skills, and abilities (KSA). Students see a rating table with three sets of behaviors that characterize each of the five attributes, and select which one of five "levels" for the attribute best characterizes the behaviors of each person on their team. This information is used to create an average rating for each person on the team for each of the five characteristics on a scale of 1 to 5. In addition, CATME provides instructors with an adjustment factor that reflects the individual student's average rating across all five categories as compared the average of the team overall. This factor can then be used to adjust team deliverable scores to individual grades. CATME has been widely used in engineering education across a range of disciplines and levels of students, with use at over 1000 institutions by nearly 6000 instructors and over 300,000 students (https://info.catme.org/about/our-user-base/).

It is important to understand the extent to which peer ratings may be influenced by unconscious or implicit bias [8]. Studies of unconscious bias have established the following situational elements as being more likely to result in unconscious bias: lack of information, time pressure, stress from competing tasks [9]. A recent review of bias in peer ratings was conducted by Stonewall et al. [10], focusing on gender and race, with acknowledgement of studies on wealth, social styles [11], friendship [11], and native/non-native English speakers. In teams of first year engineering students, women were rated lower for having related knowledge, skills, and abilities [12]. Gender and personality type were found to influence peer ratings among sophomore mechanical engineering students at a university in Canada [13]. In a large study of students at two Australian universities across six case studies, Tucker [14] found no evidence of gender bias in peer ratings for contributions to team assignments, with marginally higher ratings of women. Among first-year students in Chile, gender was not found to influence peer ratings, but an effect of "region" was found [15]. In a study with engineering students in capstone design, the predominant form of bias identified was friendship bias [16]; however, in a first-semester course with randomly assigned teams, this factor is not expected to be significant. Earlier studies from Layton and Ohland [17] and Kaufman et al. [18] found no gender bias among peer ratings by sophomore and junior engineering students, but did find rating differences based on race/ethnicity posited to be due to a variety of potential reasons including racial bias and lower actual contributions (based on lower grades among minority students). Thus, the literature is mixed regarding peer rating differences on the basis of gender and race/ethnicity.

There have been fewer studies on peer ratings for international students. Wei et al. [19] found "significant differences in peer rating behavior among international vs. domestic students in three CATME dimensions: contributing to team's work, interacting with teammates, and expecting quality." (p. 3) However, treating international students as a monolith is likely overly simplistic due to wide differences among cultures. Wei et al.'s [19] findings were situated within Hofstede's cultural dimensions and focused on individualism vs collectivism, with the international students in their study seemingly predominated by students from China, India, and South Korea. The teams may also have been predominated by first year students; the research states that these three to four person teams were situated within an introductory engineering course. Unconscious bias during peer ratings is perhaps more likely to occur among first-year

students who may have previously had limited interactions with diverse groups, in particular international students.

Although there is little literature about peer ratings of international students, documentation of racism and prejudice felt by international students suggests that peer ratings may be subject to bias. For example, prejudice and racism experiences have been felt by Asian students' at Canadian universities [20,21] and Saudi Arabian students at American universities [22,23]. A qualitative study specifically on Saudi Arabian students working on team projects in engineering / technology courses at a U.S. institution found "encompassing effects of language deficiencies, lack of understanding of the local culture, difficulties adjusting to a mixed gender environment, neo-racism, and incompatibilities in the held values" [24].

Cultural differences are likely to impact how students work in teams, including their preferences and style. Two widely used cultural comparison systems are Hofstede's [25] and the Global Leadership and Organizational Behavior Effectiveness (GLOBE) system [26]. Hofstede developed a 6-dimension system to characterize national culture on a 0 to 100 scale [25], while the GLOBE project has 9 dimensions on a 1 to 7 scale [26]. The GLOBE system also distinguishes between societal practices and societal values. These measures show some similarities among Anglo countries (USA, Canada, Australia, New Zealand, UK) and significant differences among other countries. Table 1 summarizes dimensions and values for a few countries, highlighting outliers. Students may not individually conform to their country averages, perhaps particularly likely among students who choose to study abroad. In addition, differences in survey response styles [27] might impact self and/or peer ratings in CATME.

Table 1. Examples of cultural measures from Hofstede (H) and GLOBE (G) societal practices

Dimension	Definition	USA	Mexico	Saudi Arabia	Kuwait	China
Power distance [H]	Endorses authority, power	40	81	95	90	80
Power distance [G]	differences, status privileges	4.88	5.22	NA	5.12	5.04
Masculinity [H]	More gendered	62	69	60	40	66
Gender egalitarianism [G]	Minimize gender inequality	3.34	3.64	NA	2.58	3.05
Uncertainty avoidance [H]	Seek formal procedures	46	82	80	80	30
Uncertainty avoidance [G]		4.15	4.18	NA	4.21	4.94
Individualism [H]	People feel independent	91	30	25	25	20
Performance orientation [G]	Collective encourages and rewards group members for excellence	4.49	4.10	NA	3.95	4.45
Assertiveness [G]	Individuals should be assertive / confrontational with others	4.55	4.45	NA	3.63	3.76
In-Group collectivism [G]	Individuals express pride, loyalty, & cohesiveness in organizations	4.25	5.71	NA	5.80	5.80
Team oriented leadership [G]	Team building & common purpose	5.80	5.74	NA	5.89	5.57

[H] https://www.hofstede-insights.com/country-comparison/

[G] https://globeproject.com/study_2004_2007?page_id=data#data

NA = Saudi Arabia not among the 24 countries surveyed in the GLOBE project

These global cultural frameworks provide insights regarding teamwork and caution against singular analysis of "international students". Previous research on teamwork has been grounded in global cultural dimensions [e.g. 28-30]. Based on this research, one can posit potential scenarios. For example, the uncertainty avoidance characteristic of Hofstede may indicate that

Middle Eastern students will be less comfortable engaging in an open-ended design project or overly rely on objective measures of project performance compared to US peers. The power distance characteristic may indicate that US students are comfortable with shared leadership versus a student from Saudi Arabia who may prefer that a single leader is identified and given power. The assertiveness dimension indicates that US students may be more assertive in team decisions, while students from Kuwait and China may not openly disagree with their teammates. Hispanic students may differ from majority white peers from predominantly European backgrounds due to different levels of acculturation [e.g. 31,32].

A few studies have explored the impact of international students on teams. In a first-year engineering class, Beigpourian et al. [33] found that "teams with no international students had higher psychological safety than teams with 50-67% international students." (pg. 2) Teams were composed of 3 to 4 students, and the nationalities of the international students were not provided. The authors described communication challenges as likely occurring on teams with international students, and ultimately recommended having a single international student per team (only 2 among 409 teams had 100% international students, so the sample was too small to draw conclusions about teams fully comprised of international students). Teams of first-year students that included international students were found to be "less happy" based on lower team satisfaction ratings [12]. In contrast, in a detailed interview-based study focusing on a team of four first-year male students, including one from Japan and an Asian American, "the students focused on getting assignments and projects completed as the particular value of working in diverse teams" and "felt they were more homogeneous than different" [34, p. 12, 17]. Another study [35] found lower team cohesiveness, higher conflict, and lower team satisfaction on teams in a first-semester, first-year course where half or more of the students were international; these were teams of three to four students where the non-US students were predominantly from China, India, and South Korea. The study found difference among students' perceptions depending on country and whether or not they had previously taken courses taught in English. The authors concluded, "language proficiency might have some effect on the perception of team dynamics but it is not the main factor driving it." [35, p. 6]. Relatedly, Rodriguez-Simmonds et al. [36] reported "The more similarly oriented the students are, the easier they are able to come together inclusively, working smoothly towards a common goal." This same study reported that "unwillingness to take action to support diverse groups [ ] increased" (p. 6) among the first-year students between the pre- and post- surveys. In summary, a number of studies have identified challenges that may occur when students work in diverse teams that include international students.

Research Question

This research examined whether there was evidence of implicit bias toward different demographic groups in the peer ratings associated with a team design project in a first-year engineering course.

Research Methodology: Case Study

The methodology used to address the research question was a case study. The research was situated within a retrospective analysis using data from a single course taught by the author

(2010-2019). The case study site was an *Introduction to Civil Engineering* course taught at a large, public institution; in 2012-2016 the course was co-taught with *Introduction to Architectural Engineering*. The course was required for entering first-year students majoring in these disciplines. The learning goals of the course emphasized describing these engineering disciplines and the role of ethics, sustainability, and professional licensure.

A multi-week team project to design and analyze a bridge was embedded in the course. Students were assembled into teams of 4 to 5 students, facilitated via the Team-Maker software [6]. The teaming process was structured to group students based on: times in their schedules to meet outside of class, avoid isolating female students (i.e. teams never included a single female student), similar reported commitment level and leadership preference, and differences in writing skills, software skills, leadership role, and big picture/detail-orientation. The majority of the ~600 students in the course were US residents, Caucasian non-Hispanic, and male. The course also enrolled a variety of international students, primarily from the Middle East (Saudi Arabia, Kuwait, UAE, Bahrain predominating), with a few from China, India, Nordic countries, and Latin America (Colombia, Brazil). Student nationality and race/ethnicity were not considered as grouping parameters in the team formation process.

The students were explicitly introduced to the importance of teamwork and the elements that comprise good teamwork in a lecture. When introducing the project, students were told that they would be rating their peers at the end of the project, including a brief introduction to the elements they would be rating in CATME.

The project required individuals to create a personal bridge design using software (Westpoint Bridge Designer 2010-2016, Bridge Designer 2017-2019). The team selected one of those bridges to advance to a class wide competition, or could work together to design an optimized bridge. The bridges were expected to represent an optimal combination of safety (low deflection), low cost, environmental impact, and social factors. The relative importance of these factors varied each year, illustrating the trade-offs and stakeholder-specific nature of public infrastructure design. The project typically spanned 3 to 5 weeks of class. Each team generated a single written report. The report requirements also evolved over time, such as the inclusion of an explicit discussion of sustainability and ethical issues starting in 2017. The overall project grade was based on the technical quality of the bridge itself (about half) and the quality of the write-up (about half).

After the project was completed, students rated themselves and their peers in CATME. Students were awarded points for completing CATME, but typically ~8% of the students failed to complete these ratings. When CATME added the calibration option, students were required to complete this calibration exercise. The calibration exercise presented descriptions for individuals on a hypothetical team; students then rated each team member using the CATME categories. In the first year, students were required to calibrate until sufficient agreement with correct ratings were achieved; a few students complained they had completed the calibration exercise 3 to 5 times and were unable to pass it successfully. (The software would determine whether or not sufficient similarity between student answers and the correct ratings were achieved. The calibration process in CATME has since been changed.) In subsequent years,

students were required to complete the calibration exercise at least once before rating their peers. Students can select 'reveal answers' to see how their ratings compare to the intended responses.

Research Methods

The data comprising the study included students' peer and self ratings in CATME from 2010 to 2019. Combining these years increased the amount of data in the study (total of 607 students); potential time differences were not analyzed. Each year, 87-100% of the students completed the CATME ratings (median 92%). CATME raw scores were downloaded, including the scores for the five subscales and the overall adjustment factors. The raw adjustment factors are not capped at 1.05 nor are adjustment factors of 0.95 to 1.0 rounded to 1.0. CATME outputs the data sorted per student rated. However, the individuals rating the student can be determined based on rater numbers. The author self-sorted the data so that the ratings given by particular students to other particular students were determined. An overall score given to other students averaged across the five sub-scales was computed. Demographic characteristics of the students were added. Statistical tests were conducted in Excel, including t-tests (two-tailed, heteroscedastic). Statistically significant differences were inferred when p values were below 0.05.

In addition, the graded bridge reports were examined. Each team submitted a single team report. Teams used a weighted decision matrix to select the best bridge from among the bridges designed by each individual on the team. The relative weights of the criteria were determined by the professor and established in the assignment (varied in different years of the course). Two of the rating criteria were objective: cost (computed by the bridge software) and technical (deflection visible during the animation phase of the load test). The environmental impact category was also largely objective; in later semesters of the course, the assignment specifications detailed the number of points lost for various bridge design attributes. The social factor (including aesthetics) was the only criteria that was subjective. Thus the bridge selected by the team as the best is considered to be an indicator of the quality of the design generated by team members, and would not be a "subjective" rating that could experience significant bias, unlike the peer ratings.

Limitations. A key assumption was that averaged across all ~600 students in the dataset there would not be differences in actual teamwork contributions among students in the different demographic groups being explored (gender, race/ethnicity, nationality). Therefore, statistical differences in the peer ratings might indicate bias. In some cases the number of students in these groups was somewhat small. Groups not analyzed uniquely due to very low numbers included: African Americans/Blacks, Asian Americans, other underrepresented race/ethnic groups, and international students outside the Middle Eastern and Chinese groupings. Language / culture differences may have led students to interpret the CATME rating categories differently. The struggles of some students to "calibrate" using the exercises embedded in CATME points to this as a consideration. Students were experiencing a range of environments and activities in addition to the team project that may have impacted the results. For example, a majority of the first-year civil engineering students were also taking a first-year projects course with a semester-long team project and associated teaming training (including social styles and team growth activities).

Results and Discussion

This results section explores the data with respect to potential differences in peer ratings, and then describes the quality of bridge designs within the team reports as a more objective measure of contribution quality. The results are separated into three demographic comparisons: gender, Hispanic/Latinx, and international students. This section concludes with some overall comments.

Gender

There was no evidence of gender bias found among the student ratings taken as a whole (Table 2). Female students were rated slightly lower in having relevant knowledge, skills, and abilities (median 4.4 male vs. 4.25 female), similar to the finding by [12]. The general lack of gender bias in adjustment factors resulting from peer ratings agrees with literature where gender did not impact students' peer ratings [14,15].

Table 2. Average and standard deviation of student ratings from peers and self, 2010-2019

Gender of student rated	n	Adjustment factor with self	Adjustment factor without self	Contributing (C)	Interacting (I)	Keeping team on track (K)	Expecting high quality (E)	Having knowledge, skills, abilities (H)
Male	438	.996 ± .111	.990 ± .125	4.11 ± .68	4.19 ± .59	4.07±.65	4.17 ± .56	4.30 ± .55
Female	169	1.000 ± .116	.990 ± .133	4.07 ± .68	4.16 ± .62	4.04±.66	4.12 ± .58	4.19 ± .56
p value		*.70*	*.97*	*.49*	*.62*	*.59*	*.41*	*.034*

However, gender bias may be present among the scores given by some sub-sets of students (Table 3). Specifically, Middle Eastern male students rated female students lower overall than male students, based on the average across the 5 CATME rating categories. This appears congruent with the cultural attributes of lower gender egalitarianism in Kuwait versus USA determined by the GLOBE study (see Table 1). Majority male students also rated female students slightly lower than male students on average overall.

Table 3. Average and standard deviation in the ratings given to peers by sub-groups of students (excluding self ratings)

Rater Group	Rating	C	I	K	E	H	Overall
Male MidEast	Males	4.2 ± 1.0	4.2 ± 1.0	4.0 ± 1.1	4.2 ± 1.0	4.3 ± 1.0	4.2 ± 1.0
(n=41)	Females	4.0 ± 1.1	4.0 ± 1.1	3.9 ± 1.1	4.0 ± 1.1	3.9 ± 1.1*	3.9 ± 1.1*
Male Hispanic	Males	4.2 ± 1.0	4.2 ± .8	4.1 ± 1.0	4.0 ± .9	4.2 ± .8	4.1 ± .9
(n=56)	Females	4.4 ± .8	4.3 ± .9	4.3 ± .7	4.4 ± .9	4.3 ± .7	4.3 ± .8*
Male majority	Males	4.1 ± 1.0	4.2 ± 0.9	4.1 ± 1.0	4.2 ± .9	4.3 ± .9	4.2 ± .9
(n=312)	Females	4.0 ± 1.0	4.2 ± 1.0	4.1 ± 1.0	4.2 ± .9	4.2 ± .9	4.1 ± 1.0*
Female majority	Males	3.8 ± 1.1	4.0 ± 1.0	3.7 ± 1.2	4.0 ± .9	4.3 ± .9	4.0 ± 1.0
(n=112)	Females	3.9 ± 1.0	4.1 ± 1.0	3.9 ± 1.1	4.0 ± .9	4.2 ± .9	4.0 ± 1.0
Female Mideast	Males	4.1 ± 1.1	4.2 ± 1.0	4.2 ± 1.0	4.2 ± .9	4.3 ± 1.0	4.2 ± 1.0
(n=33)	Females	4.1 ± 1.1	4.3 ± .8	4.0 ± 1.1	4.0 ± 1.0	4.0 ± 1.0	4.1 ± 1.0
Female Hispanic	Males	3.8 ± 1.3	3.9 ± 1.1	3.7 ± 1.1	3.9 ± 1.1	4.1 ± 1.1	3.9 ± 1.2
(n=16)	Females	3.6 ± 1.4	3.9 ± 1.3	3.6 ± 1.4	3.4 ± 1.3	3.8 ± 1.2	3.7 ± 1.3

* $p < .05$ comparing ratings given to female versus male students

In contrast, male Hispanic students rated females higher than male students on average overall. Due to low numbers in some groups, apparent differences based on the averages were not statistically significant (e.g. Hispanic female students' slightly higher male than female ratings overall, p = .18; Middle Eastern female students rated males somewhat higher in H = having KSA, p=.17).

In terms of the best bridges selected by the teams, there were 82 teams that included female students (most commonly 2 per team; 17 teams with 3 females) and 33 teams where the best bridge was designed by a female student (40%). This is similar to the percentage that would be predicted statistically given an average of 2 female students per team of 5 students (40% probability of a female student designing the best bridge). Thus, the objective rating of female students' technical contributions seem similar to male students.

Hispanic/Latinx Students

Hispanic/Latinx students on average were not rated differently than majority peers (Table 4), with the exception of a slightly higher 'having knowledge, skills, abilities'. Hispanic students designed the best bridge 20 times out of 63 teams with Hispanic students (31%). This is similar to the probability one would expect statistically, given an average of 1.4 Hispanic students per team (most commonly the teams included a single Hispanic student, about one-third of the teams included 2 Hispanic students, and three teams had 3 Hispanic students). There is no evidence of bias against Hispanic/Latinx students in the peer ratings.

Table 4. Peer ratings of demographic groups: Average and standard deviation

Group	n	Adjustment factor with self rating	Adjustment factor without self rating	Contributing (C)	Interacting (I)	Keeping team on track (K)	Expecting high quality (E)	Having knowledge, skills, abilities (H)
Majority	424	1.00 ± .12	1.00 ± .13	4.2 ± .7	4.2 ± .6	4.1 ± .6	4.2 ± .6	4.3 ± .6
Hispanic	75	1.01 ± .07	1.01 ± .08	4.2 ± .5	4.3 ± .5	4.2 ± .5	4.3 ± .4	4.4 ± .4*
Middle East	81	0.96 ± .12*	.94 ± .14**	3.9 ± .7**	4.0 ± .6*	3.9±.7*	4.0 ± .6*	4.1 ± .6*
Chinese	27	0.96 ± .11*	.93 ± .12*	3.6 ± .8**	3.9 ± .6*	3.6±.7*	3.9 ±. 5*	4.1 ± .5

In 2-tailed t-test compared to majority students: * $p < .05$, ** $p < .001$

International Students

Two groups of international students were explored, given the sufficiently high number enrolled in the course: Middle Eastern (primarily residents of Saudi Arabia and Kuwait) and Chinese; results summarized in Table 4. It is evident that on average international students were viewed by their peers as lower contributors, based on statistically lower adjustment factors without self ratings. Looking at the specific rating categories, Chinese students appear to have lower ratings than Middle Eastern students, but these differences are not statistically significant (p 0.13, 0.34, 0.15, 0.23, 0.96 for contributing, interacting, keeping team on track, expecting high quality, and having KSA, respectively). The results are likely confounded somewhat by gender (data not shown). When only male students from the Middle East were compared to Chinese males, statistically significant differences were found in contributing (p .03) and interacting (p .04), due to Chinese males have lower ratings than Chinese females.

The team bridge reports revealed that among the 62 teams with Middle Eastern students, a Middle Eastern student designed the best bridge 13 times (21%). This is similar to the probability one would expect on teams of 5 students with 1 Middle Eastern student. There were 24 teams with Chinese students, and a Chinese student designed the best bridge on 3 teams (13%). Thus, Chinese students' bridges were selected as the best among the team less frequently than would be expected (20-25% expected, since generally there was 1 Chinese student on a team of 4 or 5 students). Thus, based on the technical quality of the bridge designs, the lower peer ratings for Middle Eastern students may reflect biased ratings by peers, while the lower peer ratings of Chinese students appear to have some objective justification.

It is possible that in fact Middle Eastern and Chinese students contributed less to the team project than majority peers, on average. Putting together the written report and completing the group discussion elements could be distributed unevenly among team members. For example, those with better writing skills may naturally take on these tasks. While the quality of the bridge itself created by Middle Eastern students was comparable to average students in the course, the Middle Eastern students could have failed to carry their weight in the discussion and writing portions of the team report. In particular, the ethics and sustainability discussions may have proved challenging. Coming from different cultural backgrounds these students may have had different ideas about ethics or sustainability from the U.S.-centric perspective presented in-class and likely dominant among teammates.

The instructor observed some differences among international students' behaviors and interactions during the course as a whole that may have impacted peer ratings. Middle Eastern male students tended to isolate themselves from other students in class, siting together in the back of the classroom. Many of the Middle Eastern female students wore head coverings or hajib. Many Chinese students were quiet and had limited interactions with other students in the first year course. The Middle Eastern students tended to have good spoken English skills, in contrast with greater struggles among many Chinese students to communicate orally during teamwork. These behavior and/or clothing differences may have impacted how international students were viewed by their peers. For example, a quiet student may be perceived as contributing less than those who are more assertive.

General Remarks

Overall, students were likely overly generous in rating their peers, so-called leniency bias [37,38]. One would expect the average student ratings to be closer to 3, but they were in fact about 4. For example, a rating of 3 in "contributing to the team's work" is described in CATME as: "completes a fair share of the team's work with acceptable quality, keeps commitments and completes assignments on time, and fills in for teammates when it is easy or important." Thus it seems unlikely that everyone on the team would have ratings above 3, although that was often the case.

The course includes methods to increase the accuracy of peer ratings [39], including using the behaviorally anchored rating system (CATME), familiarizing the students with the dimensions and scale by which their performance will be rated (at the start of the project during the lecture

describing teamwork and on the assignment requirements document the CATME rating scales are presented), and training for ratings (using the built-in calibration exercises in CATME). However, the results from this research point to the need to also include an explicit discussion of bias in ratings with the students.

Summary and Implications

There were not statistically significant differences overall in the adjustment factors between male and female students based on peer assessments in CATME, although the average rating for "having knowledge, skills, and abilities" was slightly lower for female students. However, sub-groups of students did rate female students differently than male students, based on the average across the 5 rating categories in CATME; male Middle Eastern and male Majority students rated female students lower while Hispanic males rated female students higher. The adjustment factors for Hispanic students did not differ from majority peers, but were lower for Middle Eastern and Chinese students.

Training in implicit bias is recommended for students in courses with projects. This aligns with longstanding recommendations for rater error training [39]. However, Rodriguez-Simmonds et al. [34] found minimal benefit to this instruction on students' attitudes. A related idea might be to discuss cultural differences across countries and how this could impact teamwork. This might be best situated within a leadership discussion, contrasting the leadership styles identified in the GLOBE study.

Regardless, instructors should not expect these biases to be eliminated from peer ratings. Implicit bias might have been a factor in the lower peer ratings given to international students, which could then impact the grades awarded to international students if instructors directly used only the CATME adjustment factors. Biased ratings may also pose a threat on small teams where Middle Eastern male students may unjustifiably down-rate female students. Therefore, instructors should consider potential student bias in peer evaluations when awarding individual grades from team projects, and avoid directly using the adjustment factors from CATME. Qualitative information provided by students in CATME on who contributed to what elements in a project will help provide context, and instructors should consider this among other information to develop a more authentic understanding of teamwork processes and contributions.

References

[1] ABET Engineering Accreditation Commission. 2018. *Criteria for Accrediting Engineering Programs*. ABET, Baltimore MD.

[2] American Society of Civil Engineers (ASCE). 2008. *Civil Engineering Body of Knowledge: Preparing the Future Civil Engineer. Second Edition*. ASCE: Reston VA.

[3] Grandin, J.M. and E.D. Hirleman. 2009. Educating Engineers as Global Citizens: A Call for Action / A Report of the National Summit Meeting on the Globalization of Engineering Education. *Online Journal for Global Engineering Education*, 4 (1), article 1. 28 pp.

[4] American Society of Civil Engineers (ASCE). 2019. *Civil Engineering Body of Knowledge: Preparing the Future Civil Engineer. Third Edition*. ASCE: Reston VA.

[5] Howe, S., L. Rosenbauer, S. Poulos. 2017. The 2015 Capstone Design Survey Results: Current Practices and Changes over Time. *International Journal of Engineering Education*, 33 (5), 1393-1421.

[6] Layton, R., M. Ohland, J.R. Pomeranz. 2007. Software for student team formation and peer evaluation: CAMTE incorporates Team-Maker. American Society for Engineering Education Annual Conference and Exposition, Paper AC 2007-1565, 5 pp.
[7] Ohland, M.W., M. L. Loughry, D.J. Woehr, L.G. Bullard, R.M. Felder, C.J. Finelli, R.A. Layton, H.R. Pomeranz, D.G. Schmucker. 2012. The comprehensive assessment of team member effectiveness: Development of a behaviorally anchored rating scale for self and peer evaluation. *Academy of Management Learning & Education*, 11 (4), 609-630.
[8] NASEM - National Academies of Sciences, Engineering, and Medicine. 2017. *Supporting Students' College Success: Assessment of Intrapersonal and Interpersonal Competences*. National Academies Press: Washington DC.
[9] Fiske, S.T. 2002. What we know now about bias and intergroup conflict, the problem of the century. *Current Directions in Psychological Science*. 11(4), 123-128.
[10] Stonewall, J.H., M. Dorneich, C. Dorius, J. Rongerude. 2018. A Review of Bias in Peer Assessment. CoNECD – The Collaborative Network for Engineering and Computing Diversity Conference, Crystal City VA, 9 pp.
[11] May, G.L., L.E. Gueldenzoph. 2006. The effect of social style on peer evaluation ratings in project teams. *Journal of Business Communication*, 43 (1), 4-20.
[12] Fowler, R. 2016. Demographic effects on student-reported satisfaction with teams and teammates in a first-year, team-based, problem-based course. American Society for Engineering Education Annual Conference and Exposition, Paper ID #15060, 13 pp.
[13] Ostafichuk, P.M., J. Sibley, A. G. d'Entremont, N. Shirzad. 2015. Gender and personality type influence in peer evaluation. American Society for Engineering Education Annual Conference and Exposition, Paper ID #12670, 15 pp.
[14] Tucker, R. 2014. Sex does not matter: gender bias and gender differences in peer assessments of contributions to group work. *Assessment & Evaluation in Higher Education*, 39 (3), 293-309.
[15] Cortazar, C., I. Hilliger. 2019. Work in progress: minority bias in beer evaluations at a freshman-level engineering cornerstone course. American Society for Engineering Education Annual Conference and Exposition, Paper ID #25818, 10 pp.
[16] Thompson, R.S. 2001. Reliability, validity, and bias in peer evaluations of self-directed interdependent work teams. American Society for Engineering Education Annual Conference & Exposition. 37 pp.
[17] Layton, R.A. and M.W. Ohland. 2000. Peer evaluations in teams of predominantly minority students. American Society for Engineering Education Annual Conference and Exposition, Session 2330, 7 pp.
[18] Kaufman, D.B., R.M. Felder, and H. Fuller. 1999. Peer ratings in cooperative learning teams. American Society for Engineering Education Annual Conference and Exposition, Session 1430, 12 pp.
[19] Wei, S., D.F. Ferguson, M.W. Ohland, B. Beigpourian. 2019. Examining the cultural influence on peer ratings of teammates between international and domestic students. American Society for Engineering Education Annual Conference and Exposition, Paper ID #25066, 15 pp.
[20] Samuel, E. 2004. Racism in peer-group interactions: South Asian students' experiences in Canadian academe. *Journal of College Student Development*, 45 (4), 407-424.
[21] Houshmand, S., L.B. Spanierman, R.W. Tafarodi. 2014. Excluded and avoided: racial microaggressions targeting Asian international students in Canada. *Cultural Diversity and Ethnic Minority Psychology*, 20 (3), 377-388.
[22] Heyn, M. E. 2013. Experiences of Male Saudi Arabian International Students in the United States (Unpublished doctoral dissertation). Western Michigan University, Kalamazoo, MI
[23] Caldwell, J.D. 2013. *Examining the experiences and adjustment challenges of Saudi Arabian students in the California State University system*. Dissertation. California State University, Fresno.
[24] McKean, R.A. 2016. *Exploration of Experiences and Perceptions of Saudi Arabian Students within a Team Project Setting at a U.S. University*. Dissertation, Western Michigan University. 320 pp.
[25] Hofstede Insights. G. Hofstede. Available online: https://www.hofstede-insights.com/ Accessed Jan. 27, 2020.
[26] GLOBE Global Leadership and Organizational Behavior Effectiveness. Available online: https://globeproject.com/study_2004_2007?page_id=data#data Accessed Jan. 27, 2020.
[27] Johnson, T.P., S. Shavitt, A.L. Holbrook. 2010. Chapter 6: Survey Response Styles Across Cultures. *Cross-Cultural Research Methods in Psychology*. Pp. 130-176. DOI: 10.1017/CBO9780511779381.008
[28] Gibson, C.B. and M.E. Zellmer-Bruhn. 2001. Metaphors and meaning: An intercultural analysis of the concept of teamwork. *Administrative Science Quarterly*, 46, 274-303.

[29] Deeks, M. 2004. Cross-cultural team working within The Cochrane Collaboration. Available online: https://training.cochrane.org/sites/training.cochrane.org/files/public/uploads/resources/downloadable_resources/English/crossculturalteamwork_000.pdf Accessed Jan. 27, 2020.
[30] Egea, K., S.-K. Kim, T. Andrews, K. Behrens. 2010. Approaches used by cross-cultural and cross-discipline students in teamwork for a first-year course in web design. Proc. 12th Australasian Computing Education Conference, Brisbane AU. 10 pp.
[31] Gomez, C. 2003. The relationship between acculturation, individualism/collectivism, and job attribute preferences for Hispanic MBAs. *Journal of Management Studies*, 40 (5), 1089-1105.
[32] Romero, E.J. 2004. Hispanic identity and acculturation: implications for management. *Cross Cultural Management,* 11(1), 62-71.
[33] Beigpourian, B., M.W. Ohland, D.F Ferguson. 2019. The influence of percentage of female or international students on the psychological safety of team. American Society for Engineering Education First Year Engineering Education Conference, Penn State University, Paper ID #28069, 8 pp.
[34] Rodriguez-Simmonds, H.E., N.S. Pearson, J.A. Rohde, K.P. Vealey, A. Kirn, A. Godwin. 2017. Forget diversity, our project is due. American Society for Engineering Education Annual Conference and Exposition, Paper ID #18887, 19 pp.
[35] Jimenez-Useche, I.C., M.W. Ohland, S.R. Hoffmann. 2015. Multicultural dynamics in first-year engineering teams in the U.S. American Society for Engineering Education Annual Conference and Exposition, Paper ID #12494, 11 pp.
[36] Rodriguez-Simmonds, H.E., N.S. Pearson, B.P. Jackson, T.C. Langus, J.C. Major, A. Kirn, A. Godwin. 2018. Interpersonal interactions that foster inclusion: building supports for diversity in engineering teams. American Society for Engineering Education Annual Conference and Exposition, Paper ID #22162, 11 pp.
[37] Baker, D.F. 2008. Peer assessment in small groups: a comparison of methods. *Journal of Management Education*, 32 (2), 183-209.
[38] Lee, C.J., C.R. Sugimoto, G. Zhang, B. Cronin. 2013. Bias in Peer Review. *Journal of the American Society for Information Science and Technology*, 64(1), 2-17.
[39] Smith, D. E. 1986. Training programs for performance appraisal: A review. *Academy of Management Review*, 11(1), 22-40.