Gender and Personality Type Influence in Peer Evaluation

Dr. Peter M Ostafichuk, University of British Columbia, Vancouver

Dr. Ostafichuk is a professor of teaching in Mechanical Engienering at the University of British Columbia in Vancouver, Canada. He teaches design and other topics in mechanical engineering, and is the Chair of First Year Engineering. He has extensive experience in Team-Based Learning (TBL), and he has authored books and papers on TBL and engineering design.

Mr. James Sibley, University of British Columbia, Vancouver
Dr. Agnes Germaine d'Entremont, University of British Columbia, Vancouver

Dr. Agnes d'Entremont is an Instructor in the Department of Mechanical Engineering. Her technical research in Orthopaedic Biomechanics is focused on joint motion and cartilage health with a particular concentration in pediatric hip disorders and MRI-based methods. Her teaching-related interests include team-based learning and the flipped classroom, as well as diversity and climate issues in engineering education.

Mr. Navid Shirzad, Biomedical Engineering Graduate Program, UBC

Gender and Personality Type Influence in Peer Evaluation

Abstract

The influence of gender and personality type on peer evaluation scores in a sophomore mechanical engineering design course were considered. Eight cohorts from 2006 to 2013 were considered using mixed-linear model statistical analysis. The results revealed a statistically significant effect of evaluatee's personality preference, as described by the Myers-Briggs Type Indicator in the Judging-Perceiving domain; students with a preference for Judging received peer evaluation scores 1.1% higher on average ($p = 0.001$). Gender was shown to have a significant influence on peer evaluation score received but only for females with a preference for Introversion (1.3% higher scores than males) and females with a preference for Sensing (1.4% higher scores than males). In comparing two peer evaluation instruments as part of a three-year crossover study, the behaviorally anchored rating scale (BARS) instrument had peer evaluation score differences based on gender and personality type that were larger (reaching 3.7%) and had greater significance compared to an allocation of points (single score) instrument. The observed influences are small but non-negligible; they were slightly smaller than those reported by others for communications, management, and marketing students in terms of gender and social style influence on peer evaluation. Overall, the results suggest there is potential for evaluator bias in peer evaluation – the effects are large enough that they might be of concern to students but not so large as to invalidate the tools. Steps should be considered to educate students about potential bias.

Introduction

Teamwork is an integral part of Engineering and Engineering Education.[1] Well-designed group and team projects can help students gain valuable teaming skills, and accrediting bodies require these skills of engineering graduates.[2,3] But teamwork is not without its problems. Social loafing and "I better do it myself, if I want an A" syndrome are part of many peoples experiences with group and teamwork.[4] A well-designed peer evaluation process can improve the student experience and lead to more powerful learning outcomes.

Peer evaluation can be used to foster a better team experience and to equitably recognize individual student's contributions to their team's success.[5] There are two common form of peer evaluation – formative and summative. Formative peer evaluation (not for marks) is used to encourage positive teaming behaviours and reform or decrease poor behaviors. Summative peer evaluations (for marks) are used to temper group grades, to ensure that students get what they deserve – highly performing students that contribute well to their team are rewarded and students that do not contribute do not benefit from the group grades (which are often higher than individual grades). When we design peer evaluations we often include different amounts of summative and formative, depending on the timing in the life cycle of the team[6] and the ultimate

purpose of the evaluation. But it can be difficult to design peer evaluation schemes that are valid, reliable, and at the same time feasible to implement.[7,8]

Two commonly used peer evaluation instruments include allocation of points (AoP) and behaviorally anchored rating scales (BARS) approaches.[7] AoP, sometimes called single score, is a zero-sum method that requires evaluators to divide a set number of points between their teammates.[9,10] If some teammates are to receive above average scores, at least one other teammate must receive a below average score. This ensures that the average score each evaluator gives their teammates is the same across the class, and the average score received by all students across the class is also the same. The scores are usually accompanied by written comments. Evaluatees typically receive their aggregate score and/or comments in an anonymous, randomly-ordered fashion at the end of the evaluation period, end of the course, or not at all. The behaviorally anchored rating scales (BARS) approach uses an evaluation rubric with predefined criteria and descriptors (anchors) for different levels of performance in those criteria.[11] This focuses the evaluation on specific factors important for team success, it helps communicate constructive team behaviors to students, and aims to reduce subjectivity and improve inter-rater reliability.[7,12,13] Enszer and Castellanos showed that the two approaches have a statistically significant positive correlation.[14]

Still, in spite of our best efforts to design fair peer evaluation schemes, some bias may still exist. In a meta-analysis of leaders and managers, Eagly et al. showed men's leadership and agentic behavior were evaluated more favorably compared to those of women;[15] effects of role congruity were suggested as the source of prejudice towards female leaders.[16] Harsh reported a gender bias amongst 290 senior undergraduates in business in which evaluators tended to give higher evaluations, both in terms of performance and leadership, to managers of the same gender;[17] females received higher overall evaluations than males, in contrast to the findings of Eagly. In peer evaluations, Ruble showed the presence of same-gender bias, with greater bias in peer evaluations conducted using an AoP approach in comparison to a non-zero-sum approach.[18] In terms of personality influence, in a study of 196 sophomore and junior business communications students, May and Gueldenzoph[19] demonstrated a dependence between peer evaluation and personality, in this case described according to social style theory. In a separate study with 144 managerial communications students, May demonstrated a bias where evaluators tended to give more favourable peer evaluation scores to teammates with the same personality type, again measured by social style.[20]

This paper will explore the influence of gender and personality type, as measured by the Myers Briggs Type Indicator (MBTI), on peer evaluation scores in the context of a sophomore team-based engineering design course. In particular, this study extends the existing body of knowledge by addressing three main research questions:

- Does the evaluatee's personality type, as measured by the MBTI, influence peer evaluation score received,

- Does the evaluatee's gender influence peer evaluation score received, and
- Are gender and personality type influences in peer evaluation score more or less pronounced with the BARS peer evaluation method compared to the AoP method

The methodology for this study is outlined in the following section, including the course context, peer evaluation instruments used, and statistical methods. The results of the statistical analyses are then provided in three parts: for all peer evaluations using gender and personality type as independent effects, for all peer evaluations with gender and personality combined, and for AoP compared to BARS in a three-year crossover study. The paper finishes with discussion of the results and conclusions.

Methodology

Course Context

This study was conducted at the University of British Columbia (UBC) in a sophomore mechanical design course (MECH 223). The course is part of the integrated Mech 2 Program.[21] The typical course enrollment is 115-125 students, and eight cohorts, from 2006 to 2013, were considered in this study. The course is delivered using the Team-Based Learning (TBL) approach,[22,23] with course-specific details regarding this TBL implementation extensively documented.[24,25,26] All students attend a common lecture section (i.e. there are approximately 120 students in the classroom at one time) and they are split into four sections for other activities, such as tutorials, team meetings with a teaching assistant, computer labs, and so on. The MECH 223 course is atypical in several respects: first, the course is seven weeks in duration and, other than a course in technical communication, students do not take other courses at the same time; second, the course is split into two parts (four weeks in January and three weeks in April, each with a separate major design project); and, third, the course is large in scope at seven credits (a typical course at UBC is three credits).

Project Teams

Following recommended practice, teams of six to seven students (20 teams each year) were instructor-formed[27] in order to maximize diversity,[28,29] and to minimize previously established subgroups.[22] Prior to the course, students completed an abbreviated version of the Myers-Briggs Type Indicator (MBTI) online through the TypeFocus tool (http://www.typefocus.com). A mandatory course intake questionnaire then collected each student's MBTI preferences as well as self-reported ability with hand skills, software skills, communication skills, and team skills. The above information was combined with GPA from previous courses to form teams that were heterogeneous across all personality, skill, and GPA criteria. With the exception of the Feeling preference, the remaining seven MBTI preferences were represented by at least two members on each team; in most years there were not enough students reporting a preference for Feeling to ensure they were represented on all teams in light of other team formation constraints. The same project teams were maintained for the course duration, including the January and April sessions.

Peer Evaluation Instruments

As part of course requirements, students completed six mandatory peer evaluations using the iPeer online software tool (http://ipeer.ctlt.ubc.ca/). From 2006 to 2010, evaluations were completed using an AoP method exclusively. For these evaluations, students distributed an average of 100 points per other team member; students could reward above average performance with a score above 100, but that required lowering the evaluation scores, on average, for other team members. From 2011 to 2013, teams completed three consecutive evaluations using an AoP method and three consecutive evaluations using a BARS method; half of the class used AoP in January and BARS in April, and the other half used the reverse (i.e. a crossover study). In the BARS approach, students evaluated their teammates on four criteria:

- Communication: the individual communicates effectively with other team members
- Initiative: the individual displays initiative and contributes to team management and goal setting
- Responsibility: the individual assumes responsibility for own work, and participates equitably
- Professional behavior: the individual arrives to team meetings and class on time; completes work to a professional standard

Evaluations were completed using four levels of mastery (below expectations, marginal, meets expectations, and exceeds expectations). Before the first evaluation, teams worked through a scheduled activity in which they defined for their team the specific behaviors and evidence they would look for to assign ratings in each of the categories. The raw numerical scores for the completed BARS evaluation were normalized on an evaluator-by-evaluator basis to an average of 100 to match the AoP evaluations (i.e. each evaluator's average score given was normalized to 100). Both forms of evaluation required the students to provide comments to justify the scores they assigned. After each evaluation, students received aggregate scores and anonymous, randomly-ordered comments from their teammates. For each student, the average peer evaluation score at the end of the term was multiplied against their team's net grade in order to determine the individual portion of the team grade recorded for that student.

Subjects

Data were drawn from eight cohorts, from 2006 to 2013 (n = 975). The majority of students (n = 929) completed the TypeFocus inventory and reported their MBTI preferences. The gender distribution and distribution of reported MBTI preferences is shown in Table 1. The value of n shown for each cohort is the total number of students registered (summing the numbers for MBTI preference pairs in the table yields a slightly lower value since not all students reported their MBTI). The distribution of MBTI by gender is shown for all years in Table 2.

Table 1: Distribution of MBTI Types by Cohort. I: Introversion, E: Extraversion, S: Sensing, N:iNtuition, T: Thinking, F: Feeling, J: Judging, P: Perceiving.

Cohort	N	Gender		MBTI							
		M	F	I	E	S	N	T	F	J	P
2013	121	91	30	66	54	64	56	92	28	82	38
2012	125	103	22	56	68	50	74	88	36	67	57
2011	131	113	18	68	61	53	76	99	30	67	62
2010	122	97	25	65	50	55	60	99	16	68	48
2009	128	113	15	64	56	56	64	99	21	79	41
2008	117	98	19	48	48	35	61	77	19	53	43
2007	118	110	8	57	55	45	67	86	26	66	46
2006	113	101	12	67	46	50	63	92	21	67	4
All	975	826	149	491	438	408	521	732	197	549	380

Table 2: Distribution of Gender-MBTI Type Combinations, All Years. I: Introversion, E: Extraversion, S: Sensing, N:iNtuition, T: Thinking, F: Feeling, J: Judging, P: Perceiving.

Gender	MBTI							
	I	E	S	N	T	F	J	P
M	419	364	346	437	635	148	451	332
F	72	74	62	84	97	49	98	48

Final exam grades were used to represent individual accomplishment in the formal course material and to act as a proxy for competency in the subject.

Statistical Methods

Statistical analyses were performed using mixed-linear models. Gender and MBTI domain were included as separate random effects in the first analysis. Further analysis used combinations of gender (male or female) and each MBTI domain as random effects (e.g. male and introversion, male and extroversion, female and introversion, female and extroversion, and so on). Exam grades were centered individually for each class year to remove year-to-year effects, such as changes in exam difficulty or grader influence. Peer evaluation scores were centered over all years (as evaluation scores for each team, and therefore class year, are naturally centered). Year was added to models as a fixed effect, a random effect, and both fixed and random effects. This statistical modelling allowed for missing data; therefore, students without TypeFocus results were included in the analysis – they contributed to the mean results where specific observations were missing.

A Bayesian information criterion (BIC) was used to determine optimal models (where lower BIC indicates a better model) and a 5% significance level was used. All statistical analysis was performed using STATA (StataCorp, College Station, TX).

Results

Gender and Personality Type Considered as Separate Factors

Peer evaluation score was examined with respect to exam grade, gender, and MBTI domains (Table 3). For this analysis, peer evaluation score was based on the AoP method for 2006 to 2010 and was based on the average of the AoP and BARS methods for 2011 to 2013. As shown in Table 3, a statistically significant positive correlation between exam grade and peer evaluation was observed. Based on BIC, year was not included in any of the best models selected. In comparing each gender or MBTI domain type to its dual (e.g. male to female, introversion to extraversion, etc.), the only statistically significant difference was found in the Judging-Perceiving domain. This effect was independent of the effect of exam grade on peer evaluation score. Those with a preference for Judging (characterized by a planned and organized approach and a preference to make a decision and move on) were found to receive peer evaluation scores 1.07% higher on average ($p = 0.001$) than those with a preference for Perceiving (characterized by flexibility, spontaneity, and holding off making a decision in order to collect more information).

Table 3: Examination of influence of exam grades, gender, and MBTI domains on peer evaluation score.

	Peer Evaluation Score (2006-2013)	
	Slope of peer evaluation score	p-value
Exam grade	**0.090**	**<0.001**
	Difference in peer evaluation score	p-value
Male to female	-0.465	0.27
Introversion to Extraversion	-0.045	0.88
Sensing to iNtuition	0.196	0.53
Thinking to Feeling	-0.596	0.12
Judging to Perceiving	**1.068**	**0.001**

When the peer evaluation scores from the BARS method (2011 to 2013) were removed – leaving six AoP evaluations per cohort from 2006 to 2010 and three AoP evaluations per cohort from 2011 to 2013 – a similar effect was observed where those with a preference for Judging received peer evaluation scores 0.85% higher ($p = 0.005$). Differences between AoP and BARS are considered further at the end of this section.

Gender and Personality Type Considered in Combination

The influences on peer evaluation scores received were further examined by considering combinations of gender and MBTI in the analysis. For example, using the Introversion-Extraversion domain, peer evaluation scores were compared for the four possible combinations: male-Introversion, male-Extraversion, female-Introversion, and female-Extraversion. These results for the various combinations are shown in Table 4 through Table 7 for four MBTI domains, respectively. They reveal that the effects are more complicated than suggested in Table 3. The diagonals represent self-comparisons and are greyed out; the corners of the diagonal have been omitted for compactness. Cells below the diagonal (blanks) are equal and opposite to corresponding cells above the diagonal.

For the Introversion/Extraversion domain (Table 4), a statistically significant difference was observed between females with a preference for Introversion and males (both with preferences for Introversion and Extraversion). Females with a preference for Introversion received peer evaluation scores between 1.2% and 1.4% higher than males. This effect is not revealed in the single parameter comparisons of Table 3. No other statistically significant differences were observed for other combinations involving Introversion and Extraversion.

Table 4: Comparison of peer evaluation scores for gender/IE combinations. Exam marks included as a random effect: slope of peer evaluation with exam grade = 0.088 (p-value < 0.001).

	Relative to Male-Extraversion		Relative to Female-Introversion		Relative to Female-Extraversion	
	Difference	p-value	Difference	p-value	Difference	p-value
Male-Introversion	-0.169	0.61	**-1.364**	**0.02**	-0.108	0.85
Male-Extraversion			**-1.195**	**0.05**	0.060	0.92
Female-Introversion					1.256	0.10

Similarly, for the Sensing/iNtuition domain (Table 5), a statistically significant difference was observed between females with a preference for Sensing and males (both with preferences for Sensing and iNtuition). Females with a Sensing preference received peer evaluation scores between 1.3% and 1.5% higher than males, again not revealed in Table 3. No other statistically significant differences were observed for other combinations. Although the peer evaluation score difference between females with Sensing and iNtuition preferences were of a similar magnitude and nearly statistically significant ($p = 0.07$), the sample size in each group was small (62 and 84, respectively) which may be limiting the statistical power of this comparison.

Table 5: Comparison of peer evaluation scores for gender/SN combinations. Exam marks included as a random effect: slope of peer evaluation with exam grade = 0.087 (p-value < 0.001).

	Relative to Male-iNtuition		Relative to Female-Sensing		Relative to Female-iNtuition	
	Difference	p-value	Difference	p-value	Difference	p-value
Male-Sensing	0.195	0.56	**-1.343**	**0.04**	0.053	0.93
Male-iNtuition			**-1.538**	**0.02**	-0.142	0.80
Female-Sensing					1.397	0.07

No statistically significant differences were observed for any of the combinations involving Thinking and Feeling (Table 6, included for completeness).

Table 6: Comparison of peer evaluation scores for gender/TF combinations. Exam marks included as a random effect: slope of peer evaluation with exam grade = 0.089 (p-value < 0.001).

	Relative to Male-Feeling		Relative to Female-Thinking		Relative to Female-Feeling	
	Difference	p-value	Difference	p-value	Difference	p-value
Male-Thinking	-0.544	0.20	-0.773	0.13	-0.715	0.30
Male-Feeling			-0.229	0.71	-0.171	0.82
Female-Thinking					0.058	0.94

Table 3 showed statistically significant differences for the Judging/Perceiving domain. In considering the gender-MBTI combinations (Table 7), statistically significant differences were observed for three of four combinations involving both Judging and Perceiving: male Judging to male Perceiving, female Judging to female Perceiving, and male Perceiving to female Judging. Those with a preference for Judging received peer evaluation scores between 0.9% and 2.1% higher on average. The combination of male Judging to female Perceiving did not show a statistically significant difference. Further supporting the Judging-Perceiving effect, statistically significant differences were not observed for male-female comparisons when both groups had the same personality preference (Judging or Perceiving).

Table 7: Comparison of peer evaluation scores for gender/JP combinations. Exam marks included as a random effect: slope of peer evaluation with exam grade = 0.089 (p-value < 0.001).

	Relative to Male-Perceiving		Relative to Female-Judging		Relative to Female-Perceiving	
	Difference	p-value	Difference	p-value	Difference	p-value
Male-Judging	**0.867**	**0.01**	-0.979	0.058	1.142	0.10
Male-Perceiving			**-1.847**	**0.001**	0.273	0.70
Female-Judging					**2.121**	**0.009**

Comparison of Peer Evaluation Instruments

For the comparison between AoP and BARS, data were analyzed for cases where the same students completed evaluations using both instruments (that is, three cohorts between 2011-2013, n = 377). For each type of evaluation, gender, each of four MBTI domains, and exam grade were included in the statistical model. Peer evaluation scores for both instruments correlated with exam grade, with BARS showing a stronger correlation (greater slope) and a smaller p-value. Similar to Table 3, the only statistically significant difference observed was between students with a preference for Judging and Perceiving (Judging received higher peer evaluation scores); however, this effect was only observed for the BARS evaluations and not for AoP (Table 8).

Table 8: Comparison between AoP and BARS peer evaluation instruments for years when they were both used equally.

	AoP (2011-2013)		BARS (2011-2013)	
	Slope of peer evaluation score	p-value	Slope of peer evaluation score	p-value
Exam grade	**0.064**	**0.005**	**0.144**	**<0.001**
	Difference in peer evaluation score	p-value	Difference in peer evaluation score	p-value
Male to female	-0.586	0.36	-1.448	0.16
Introversion to Extraversion	0.641	0.19	-0.680	0.38
Sensing to iNtuition	-0.260	0.61	0.406	0.62
Thinking to Feeling	-0.221	0.70	-1.310	0.15
Judging to Perceiving	0.881	0.09	**1.834**	**0.03**

As shown, for AoP from 2011 to 2013, the difference between Judging and Perceiving was 0.88% (Judging higher) but was not significant (p = 0.09). As mentioned previously, using all AoP data from 2006 to 2013 yielded a difference between Judging and Perceiving of 0.85% (Judging higher) that was statistically significant (p = 0.005). Therefore, it appears a consistent Judging-Perceiving effect is present and the lack of statistical significance in Table 8 is due to insufficient statistical power with the three cohorts that that data is drawn from.

Comparison of Peer Evaluation Instruments Using Gender-Personality Combinations

We also examined the relationship between peer evaluation score and gender-MBTI pairings separately for AoP and BARS. This was done only for the years in which students used both peer evaluation instruments (2011-2013). Statistically significant differences were not observed for any of the AoP evaluations. For BARS, statistically significant differences were observed for only the two following cases: females with a Judging preference received higher peer evaluation scores than males with a Perceiving preference (3.69 percent higher, p = 0.005), and females with a Sensing preference received peer evaluation scores higher than males (3.22 percent higher than male Sensing, p = 0.04, and 3.69 percent higher than male iNtuition, p = 0.02).

Discussion

A statistically significant positive correlation between peer evaluation score received and exam grades was observed in all models. This was accounted for in the models such that observed differences based on gender and personality type were not attributable to differences in competency in the subject matter (in so far as exam scores were a true measure of competency).

Influence of Personality Type

Considering the first research question, based on the eight cohorts from 2006 to 2013, statistically significant differences in peer evaluation score received were found in terms of MBTI personality type. The most notable effect was that students with a preference for Judging received higher peer evaluation scores (1.1%, $p = 0.001$) than their colleagues with a preference for Perceiving. None of the remaining three MBTI personality type domains (Introversion/Extraversion, Sensing/iNtuition, and Thinking/Feeling), when treated as separate factors, showed any statistically significant effect on peer evaluation score.

The differences in peer evaluation scores between those with a preference for Judging and those with a preference for Perceiving can likely be attributed to how the qualities of each preference might be observed in a team setting. The primary attributes of Judging include a planned and organized approach, and a tendency to make a decision based on available information and move on. In a team, this could be interpreted, correctly or not, as decisiveness, action, leadership, and engagement. In contrast, the primary attributes of Perceiving include flexibility, spontaneity, and the tendency to delay making a premature decision in order to make sure all the information is available to make the "right" decision. This could be interpreted, correctly or not, as procrastination, a lack of initiative, or disengagement.

Influence of Gender

Turning to the second research question, while others have reported evaluation biases of up to 2.5% based on gender,[18] we did not observe any statistically significant effect that could be attributed to gender alone. However, when gender was coupled with the MBTI personality type domains, additional statistically significant effects on peer evaluation score were identified. In total, there are 24 possible comparisons between gender-personality pairs (e.g. male-Introversion compared to male-Extraversion, female-Introversion, and female-Extraversion). We examined all 24 and 7 of these showed statistically significant effects:

- Female-Introversion received higher peer evaluation scores than male-Introversion (+1.36%, $p = 0.02$) and male-Extraversion (+1.20%, $p = 0.05$)
- Female-Sensing received higher peer evaluation scores than male-Sensing (+1.34%, $p = 0.04$) and male-iNtuition (+1.54%, $p = 0.02$)
- Male-Judging received higher peer evaluation scores than male-Perceiving (+0.87%, $p = 0.01$)
- Female-Judging received higher peer evaluation scores than male-Perceiving (+1.85%, $p = 0.01$) and female-Perceiving (+2.12%, $p = 0.009$)

In all statistically significant differences involving both Judging and Perceiving types, those with a preference for Judging received higher peer evaluation scores. This echoes the results of the initial analysis in which gender and personality type were not coupled. A second trend is also revealed in the above: in all statistically significant differences involving both females and males, females received higher peer evaluation scores than males. This suggests a gender effect does indeed exist, but only under certain circumstances. This is also in contrast to the results reported by Ruble[18] where ratings were higher when the evaluator and evaluatee had the same gender. Possible reasons for the differences could be due to the unique culture or norms of the different disciplines (engineering versus communications and marketing) or programs (e.g. due to the different schools or instructors).

Comparison of Peer Evaluation Instruments

Considering the third research question, the three-year crossover study revealed differences between the BARS and AoP methods. Specifically, differences with BARS were more pronounced than differences with AoP. The BARS evaluations showed a stronger correlation between peer evaluation score and exam grades (slope 0.144, $p < 0.001$) than did the AoP evaluations (slope 0.064, $p = 0.005$). Given the approach was that of a crossover study, this suggests better performance discrimination with BARS. Considering the effects of gender and personality type, the AoP evaluations showed no statistically significant peer evaluation score differences for the three years of the crossover study. For the BARS evaluations in the same period, a statistically significant difference was observed for the Judging/Perceiving domain (Judging 1.83% higher, $p = 0.03$). Considering the full eight years of AoP evaluations, the Judging-Perceiving difference was statistically significant ($p = 0.005$) but the observed effect (0.85%) was less than half of that of BARS.

Limitations

This study was conducted over a period of eight years with almost 1000 students. There are several limitations that must be acknowledged. First, the study was conducted at a single institution and a single course, so behavioral and cultural norms that have developed at that institution and in that course have not be accounted for. In addition, women students averaged 15% of the overall population and students with a MBTI preference for Feeling averaged 20%; this limits the statistical power of the analyses but also may affect the behavioral and cultural norms that develop. In other words, the results may not be applicable to another program or discipline that tends to attract different students.

Overall

Overall, these results suggest that gender and personality type have an influence in peer evaluation. For the statistically significant effects, the differences between dissimilar groups (e.g. male to female) was roughly 1% to 2% of the peer evaluation score – small but non-negligible. Controlling team membership to place like types together (e.g. females with a preference for Judging with other females with a preference for Judging) might reduce bias, but would be highly impractical if not impossible in some cases (in addition, it would go against best

practices of forming heterogeneous teams[27]). On the other hand, educating students about these effects and training them to be aware of potential bias when completing peer evaluations is a practical approach that has been shown to work in similar situations.[20] At the very least, instructors using peer evaluation should be aware that personality type and gender potentially introduce a bias of 1% to 2% on the scores received. To most instructors, 2% is small and lost in the "noise" of grading. In this case it is important to remember that the effects are not random noise, but rather they are attributable to personal traits an individual cannot change. Imagine telling students that simply because of their personality type or gender they can expect peer evaluation grades up to 2% lower. Again, small but non-negligible.

Conclusions

Peer evaluation data was analyzed for eight cohorts of sophomore mechanical engineering design students (n = 975). Statistical analysis using mixed-linear models revealed some significant effects in all three research questions considered. Namely:

- In considering the possible influence of an evaluatee's personality type, as described by the MBTI, a significant effect was observed for the Judging-Perceiving domain. Those with a preference for Judging received peer evaluation scores 1.1% higher on average ($p = 0.001$). No effects were observed for the other MBTI domains, provided they were not coupled with gender (see next point).

- Gender alone was not found to have a statistically significant influence on peer evaluation score, but gender coupled with personality type did show significant effects. Females with a preference for Introversion received peer evaluation scores 1.3% higher than males, and females with a preference for Sensing received peer evaluation scores 1.4% higher than males.

- For the two peer evaluation instruments considered, BARS was found to have a stronger correlation to course exam grades (2.25 times the slope of peer evaluation score to exam grade) compared to AoP. In addition, the differences between Judging and Perceiving noted above were more pronounced with BARS (1.8%) than with AoP (0.9%).

The results suggest care must be exercised in the peer evaluation process and steps to educate students about potential bias should be considered.

References

[1] Ohland, M.W., M.L. Loughry, R.L. Carter, L.G. Bullard, R.M. Felder, C.J. Finelli, R. A. Layton, and D.G. Schmucker, "Developing a Peer Evaluation Instrument that is Simple, Reliable, and Valid," *Proc. ASEE Annual Conference & Exposition*, Portland, OR, June 2005.

[2] International Engineering Alliance, *Graduate Attributes and Professional Competencies*, Version 3: 21 June 2013, Available as of Feb 2, 2015 from: http://www.ieagreements.org

[3] ABET, Criteria for Accrediting Engineering Technology Programs, Baltimore, MD: Accreditation Board for Engineering and Technology (ABET), 2013, 31 pp. Available as of Feb 2, 2015 from: http://www.abet.org/etac-criteria-2014-2015/

[4] Oakley, B., D.M. Hanna, Z. Kuzmyn, and R.M. Felder, "Best Practices Involving Teamwork in the Classroom: Results from a Survey of 6435 Engineering Student Respondents," *IEEE Transactions on Education*, Vol. 50, No. 3, 266–272 (2007).

[5] Millis, B.J. and P.G.Cottell, *Cooperative Learning for Higher Education Faculty*, Oryx Press, (1998).

[6] Delson, N.J., "Increasing team motivation in engineering design courses," *International Journal of Engineering Education* 17(4-5): 359-66 (2001)

[7] Baker, D.F. "Peer Assessment in Small Groups: A Comparison of Methods." *Journal of Management Education*, Vol. 32, No. 2, 183–209 (2008).

[8] Saavedra, R. and S.K. Kwun, "Peer evaluation in self-managing work groups." *Journal of Applied Psychology*, Vol. 78, 450-462 (1993).

[9] Fellenz, M. R., 2006, "Toward fairness in assessing student groupwork: A protocol for peer evaluation of individual contributions," *Journal of Management Education*, Vol. 30, pp. 570-591.

[10] Gueldenzoph, L. E., May, G. L., 2002, "Collaborative peer evaluation: Best practices for group member assessments." *Business Communication Quarterly*, Vol. 65, pp. 9-20.

[11] Smith, P. C. and L.M. Kendall, "Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales," *Journal of Applied Psychology*, 47: 149-155 (1963).

[12] MacDonald, H.A., L.M. Sulsky, "Rating Formats and Rater Training Redux: A Context Specific Approach for Enhancing the Effectiveness of Performance Management," *Canadian Journal of Behavioural Science*, Vol. 41, No. 4, 227-240 (2009).

[13] Ohland, M.W., R.A. Layton, M.L. Loughry, and A.G. Yuhasz, "Effects of Behavioral Anchors on Peer Evaluation Reliability." *Journal of Engineering Education*, Vol. 94, No. 3., 319–326 (2005).

[14] Enszer, J.A., M. Castellanos, "A Comparison of Peer Evaluation Methods in Capstone Design," *Proc. ASEE Annual Meeting*, Atlanta, GA, June, 2013.

[15] Eagly, A.H., S.J. Karau, and M.G. Makhijani, "Gender and the Effectiveness of Leaders: A Meta-Analysis," *Psychological Bulletin*, Vol. 117, No. 1, 125-145 (1995).

[16] Eagly, A.H., and S.J. Karau, "Role Congruity Theory of Prejudice Toward Female Leaders," *Psychological Review*, Vol. 109, No. 3, 573-598 (2002).

[17] Harsh, K.L., "Gender Differences in Evaluation of Performance and Leadership Ability: Autocratic vs. Democratic Managers," *Sex Roles*, Vol. 35, Nos. 5/6, 337-361 (1996).

[18] Ruble, T.L., S.A. Hernandez, and W.J. Amadio, "A Comparison of Peer Evaluation Systems in Team-Based Learning," *Proc. of the Academy of Business Education Annual Conference*. 2004.

[19] May, G.L., L.E. Gueldenzoph, "The Effect of Social Style on Peer Evaluation Ratings in Project Teams," *Journal of Business Communication*, Vol. 43, No. 1, 4-20, January (2006).

[20] May, G.L.," The Effect of Rater Training on Reducing Social Style Bias in Peer Evaluation," *Business Communication Quarterly*, Vol. 71, No. 3, 297-313, September (2008).

[21] Ostafichuk, P.M., E.A. Croft, S.I. Green, G.S. Schajer and S.N. Rogak, "Analysis of Mech 2: An Award-Winning Second Year Mechanical Engineering Curriculum," Proc. of EE2008, Loughborough, UK, July 2008.

[22] Michaelsen, L.K., M. Sweet, M., and D.X. Parmelee , Team-Based Learning: Small Group Learning's Next Big Step. *New Directions for Teaching and Learning*, Jossey-Bass, San Francisco (2008).

[23] Sibley, J. and P.M. Ostafichuk , *Getting Started with Team-Based Learning*, Stylus, Sterling, VA (2014).

[24] Ostafichuk, P.M., Hodgson, A.J., Bartek, S., and Naylor, C., "Teaching Team Dynamics: Experiences in Second Year Mechanical Engineering Design", Proc. CDIO Conference, Montreal, QC, June, 2010.

[25] Hodgson, A.J., and P.M. Ostafichuk, "Team-Based Learning in the Design Modules of a New, Integrated, 2nd Year Curriculum at UBC," Proc. CDEN, Kananaskis, AB, July, 2005.

[26] Hodgson, A.J. and P.M. Ostafichuk, "Designing Extended Assignments for Team-Based Learning Modules," Proc. CDEN 2006, Toronto, ON, July, 2006.

[27] Brickell, J., Porter, A., Reynolds, M., and R. Cosgrove, "Assigning Students to Groups for Engineering Design Projects: A Comparison of Five Methods," *Journal of Engineering Education*, Vol. 83, Issue 3, 259-62, July (1994).

[28] Feichtner, S., and E. Davis, "Why Some Groups Fail: A Survey of Students' Experiences with Learning Groups," *The Organizational Behavior Teaching Review*, Vol. 9 No. 4, 58-73, (1984).

[29] Weimer, M., "Why Groups Fail: Student Answers," *The Teaching Professor*, Vol. 5, No. 9, November (1991).