STUDENT EVALUATION OF TEACHING IN AN ENGINEERING CLASS AND COMPARISON OFRESULTS BASED ON INSTRUCTOR GENDER

Byron Hempel, Kasi Kiehlbaugh, and Paul Blowers
University of Arizona • Tucson, AZ 85721

Student evaluations of teaching (SETs), sometimes referred to as teacher–course evaluations (TCEs), play a fundamental role in higher education. No one in the university setting spends more time with an instructor than the enrolled students, making regular and systematic feedback from students a powerful source of information for gaining insights into instruction quality.[1] SETs are becoming increasingly important for faculty promotion decisions, student course selection, and auditing purposes.[2] Student feedback can provide actionable knowledge to deans, department heads, faculty, instructional developers, teaching assistants, and students. Given well-constructed questions, students can provide valid, reliable, and useful data concerning the classroom experience. Instructors can obtain information from SETs to improve their own practices, and students can decide whether to take a course with a particular instructor. Certain classroom practices also can be appraised indirectly from SETs through year-to-year or instructor-to-instructor analysis of teaching. With the use of SETs, instructors are able to improve their teaching practices, ultimately allowing students to have the greatest opportunity for learning.[3]

There are limitations and resistance to the exclusive use of SETs for making promotion and tenure decisions. In particular, the methodology used to construct SETs is important. The evaluations need to contain robust questions and be able to provide clear, unbiased results. Of the many possible biases present in SETs,[3]–[7] this paper focuses on gender bias. Although the literature does not show clear agreement on the issue, many prior papers have discussed how instructor gender may affect SETs.[3] Much of the prior research has involved laboratory studies, which attempt to control aspects of courses or instructors that may influence student results. In these types of studies, students may read descriptions of professors and then complete SETs.[3] Students may be asked to look at photographs and then rate instructors or may be given syllabi and be supplied with gender cues and then evaluate the hypothetical teaching ability of the instructors based on the limited information.[8] In contrast, observational studies are "real life" in that they collect data from enrolled students in live classes,[9] often across sections or semesters to pool enough data to provide meaningful analyses.[10]–[12] These types of studies have become more common in recent years as automated university data collection, electronic ar-

Byron Hempel *is a Ph.D. Candidate at the University of Arizona, having received his B.S. in Chemistry at the University of Kentucky and M.S. in the Chemical and Environmental Engineering Department at the University of Arizona. Working under Dr. Paul Blowers, Byron is focusing on improving the classroom environment in higher education by working in a flipped classroom.*

Dr. Paul Blowers *is a full professor and a University Distinguished Professor at the University of Arizona. He received his B.S. in Chem. Eng. from Michigan State University in 1994 before going on to receive an M.S. and Ph.D. from UIUC in 1997 and 1999 in Chem. Eng.*

Dr. Kasi Kiehlbaugh i*s a career-track lecturer at the University of Arizona. She received her B.S. and M.S. in Chemical Engineering at the University of Arizona, afterwards receiving her Ph.D. in Chemical Engineering at UC-Berkley. She is primarily interested in incorporating research-based pedagogical techniques into the undergraduate engineering classroom, and she focuses on employing active learning techniques and utilizing collaborative learning space classrooms.*

chival of the ratings, and statistical methods for handling large data sets have become available. One such study was published by Fenn, who determined that the gender and race of the instructor are significant, with female and non-white instructors being downgraded by students within Economics and Business.[5] Having a greater understanding of SETs and the student biases that are present in those assessments will allow those using SETs to mitigate how student biases may influence promotion and tenure decisions and other reward structures. As the culture of higher education is moving toward recognizing and rewarding quality teaching at research universities, it is paramount to recognize the biases inherent in one of the main forms of evaluation.[13]

Because SETs are often the only method of instructor evaluation used at higher education institutions, recognizing biases that exist will allow departments to appropriately assess instructors.[14] However, there does not seem to be a consensus on student biases, and the literature varies among fields. For instance, Nadler et al. showed that psychology students showed an increased gender bias against female faculty, whereas law students showed a decreased gender bias.[15] The lack of identifiable trends in the literature on SETs exists for several reasons. Most of the experiments did not occur in a field setting or only involved one lecture or presentation by an instructor,[16] making it difficult to know whether the results would hold true in regular classrooms during a full semester.[17] Basow and Silberg suggested that it would be useful to look at classroom behaviors to see how those behaviors affected teaching effectiveness.[6] Spooren et al. reviewed the SET literature in 2013 and noted that many biasing factors could impact instructor evaluations, including the course discipline or sexual orientation of the instructor,[18] and that gender could play additional roles. Additionally, when a field of study is largely segregated between men and women, various sources suggest that there is a gender minority bias when students evaluate instructors of the majority gender.[19] Anderson and Smith pointed out that one would have to control for the exact content of the course, including lectures, requirements, and time of day to systematically determine whether there are gender biases.[20] To address the limitations of prior studies, we set up an educational experiment in which aspects of the classroom experience for students were as uniform as possible other than the experience level, age, and gender of the faculty members.

The goal of this study was to provide insight into how gender biases may influence SET scores within a highly controlled environment. A co-teaching model was used in which each faculty member instructed with the same teaching model and taught for the same number of lectures throughout the semester. The model closely follows one that Bullard and Felder suggested in the context of a chemical engineering curriculum.[21] By controlling multiple aspects of the course, we created a significantly more controlled environment than studies in the past. Based on past research and knowing that chemical engineering is a largely male-dominated field,[22] our hypotheses were that the following would occur in a core sophomore chemical engineering course:

1. *The female instructor would be evaluated by students as a less effective instructor through Likert scale rankings of instruction.*
2. *The female instructor would receive more negative open-ended comments than the male instructor.*

This study explores students' views of male and female instructors in chemical engineering.

METHODS

Anderson and Smith pointed out that one would have to systematically control for the exact content of the course, including lectures, requirements, and time of day, to determine whether there are gender biases.[20] We argue that we not only controlled these variables but that we also controlled the students, the classroom configuration, the difficulty level of the material, and the teaching style with high fidelity.

INSTRUCTOR PROFILES

The male instructor had taught at the university for 16 years when the experiment started, is a tenured full professor, and is a University Distinguished Professor, which is the highest teaching award given by the institution for excellence in undergraduate instruction. In the years leading up to the experiment, this instructor had used a predominantly active-based approach with partially-flipped hybrid classes in which students do pre-lecture activities (for example, watching a brief recorded lecture, reading the textbook, and/or completing a reading quiz) prior to attending class. This instructor has used some form of active-based teaching for 15 years.[23], [24]

The female instructor had taught at the university for 5 years when the experiment started, was a career-track lecturer, and won a university-wide teaching award in the second year of this experiment. Her teaching experience before teaching the experimental course included two 1-unit laboratory courses and a 2-unit freshman engineering course in which she deployed formative assessment technologies such as clickers and other web-based tools to enhance instruction.[25] She was just starting to explore active teaching techniques prior to the first experiment. Both instructors had received PhDs in chemical engineering, so their fundamental knowledge was similar. Neither instructor had previously taught the material in the course; this was a new classroom experience for both instructors. They worked together to collaboratively build an active course.

CLASSROOM DESIGN

The class chosen for this research was a sophomore-level core chemical engineering course with a starting enrollment of approximately 90 students. In terms of gender, the students were 25-30% female. The class is a 4-credit class that met three times per week for a 50-minute lecture and again for another 50 minutes in small breakout sessions with a teaching assistant or an undergraduate learning coach (also known as a preceptor) to solve homework problems that are due the following week.

In the spring of 2016, both instructors involved with the class were teaching the course for the first time and did not have materials from a prior instructor. The instructors developed the course content as the semester unfolded. Course materials included classroom pre-quizzes that were due before 36 of the 45 class meetings; the pre-quiz scores represented 10% of the total grade in the course. Attendance in class was required and made up another 10% of the total grade. The male instructor created all of the exams and the rubrics for those exams, while both instructors created homework assignments, lecture materials, and pre-lecture quizzes throughout the semester. The students did not know which instructor had created specific content or assignments for the class. On any given day, one faculty member would have created the pre-lecture quiz while the other created the lecture content. One of the professors would then deliver the content in class, while the other was present and functioned as a preceptor, answering questions from students and helping students solve the active engagement problems. The professor not lecturing would take notes of potential improvements to the lecture. Because of the consistent peer observation, feedback and collaboration on content creation, there was a rapid convergence in teaching style and presentational mode, eliminating limitations of some prior work[26] in which teaching mode was an interfering variable.[27] Spooren et al. discussed how the number of control variables included in studies, how variables were measured, and the research techniques could interact with the sample characteristics to lead to confounding results in SETs[18]. We believe that the rapid convergence of teaching styles, as measured by the COPUS tool, described later, eliminated many of the potentially confounding variables.

The spring 2016 course was broken into six discussion sections, and the 2017 course was broken into seven sections, with between nine and twenty-four students in each section. Each section met separately for one hour per week for discussion and small-group problem solving outside of the main three hour-long lectures held per week. In 2016, the discussion sections were facilitated by a teaching assistant, and in 2017, by undergraduate student preceptors. Neither instructor was involved in that portion of the class other than selecting materials that the teaching assistant or preceptors would use to aid in student learning.

In the second year of co-teaching, the instructors would often switch and deliver the materials/slides created by the other instructor in the first year. The largest variance in the classroom experience was the lecture creation process in year one when the two instructors had different foci. The female instructor was more detail oriented in some aspects of the delivery while the male instructor was bigger-picture oriented when content was created. In year two, by presenting each other's slides with little or no modification, the differences in instructor experience and preference in content presentation were removed. Both instructors adapted simultaneously to the new co-teaching model. The instructors developed ground rules where both instructors were comfortable taking control of the class to lead a short section if they felt the other instructor was missing details or was failing to notice a student misconception when students were struggling with content. This relinquishing of control by one instructor to the other was equally common for both the male and female instructor.

Both brought different skills to the classroom. The male instructor had never used the clicker/voting technologies that the female instructor had been using and capitalized on the instructional affordances that the technology brought in year one. The female instructor had been using active methods in the small seminar setting, but, unlike the male instructor, had never taught a 80+ person class that met three times a week. Personal reflections and observations determined that both instructors were convergent in teaching style.

STATISTICAL ANALYSIS

The institution where this work was conducted is a large Research I institution and is a member of the Association of American Universities. SETs (termed TCEs locally) are conducted automatically in the last four weeks of the semester in an online format. To maintain student confidentiality, the TCEs at the study's R1 institution do not report the individualized demographics of the students, and there is no way to track responses back to students, even through de-identification. To encourage students to complete the assessment, a small number of bonus points (representing ~1% of the grade in the course) were offered if the class response rate was above 90%. In the years included in the study, the response rates were high, with >95% percent responses each year. Of the 14 Likert questions prompted on the SET, the following five questions were analyzed for this study:

1. *"What is your overall rating of this instructor's teaching effectiveness?" with a value of 5 being "almost always effective", 4 "usually effective", 3 "sometimes effective", 2 "rarely effective" and 1 "almost never effective"*
2. *"What is your overall rating of this team's teaching effectiveness?" with a value of 5 being "almost always effective", 4 "usually effective", 3 "sometimes effective", 2 "rarely effective" and 1 "almost never effective"*

3. *"What is your overall rating of this instructor compared with other instructors you have had?" with a value of 5 being "one of the most effective", 4 "more effective than usual", 3 "about as effective as usual", 2 "less effective than usual" and 1 "one of the least effective"*

4. *"What is your overall rating of this team's teaching effectiveness compared with other teaching teams?" with a value of 5 being "almost always effective", 4 "usually effective", 3 "sometimes effective", 2 "rarely effective" and 1 "almost never effective"*

5. *"I was treated with respect in this course" with a value of 5 being "strongly agree", 4 "agree", 3 "uncertain", 2 "disagree" and 1 "strongly disagree"*

In addition to the TCEs from 2016 and 2017, data were collected on the previous instructors' ratings for the course for comparison to the study instructors. Because the scope of the paper does not extend into historical performance evaluation and there were significant differences in response rate and variance in the years prior to the study, no statistical t-tests were performed on the historical data. The two categories collected for historical comparison were the "overall rating of the instructor's effectiveness" and the "comparison to other instructors" found in Table 1.

A standard Student's t-test with equal variance was used to compare Likert data. Zimmerman concluded that optimum protection from Type I errors is assured by using the Welch test whenever sample sizes are unequal.[28] Because the sample sizes were identical within years and very close between years, the Welch test was unnecessary, as no corrections for heteroscedasticity were necessary.[28],[29] Because the samples show no significant differences in variance, neither the unequal variance t-test nor the Mann–Whitney U test was used.[30] For five-point Likert items, de Winter and Dodou[31] found that t-tests and Mann–Whitney–Wilcoxon (MWW) tests had no Type 1 errors above 3% in a randomized population, suggesting that t-tests and MWW tests have similar power. Norman also concluded that parametric statistics can be used with Likert data, "with small sample sizes, with unequal variances, and with non-normal distributions, with no fear of coming to the wrong conclusion."[32] We provide various parameters to avoid misconceptions, such as assuming that ordinal Likert data yields perfect interval responses between numbers. A normal alpha value of 0.05 was used; because the dependent variables were broken down into various categories, the Bonferroni alpha value of 0.01 was also used.[33],[34]

TABLE 1
Comparison of Likert data for past instructors and the study instructors

	Likert Topic:	Instructor Effectiveness		Compared with Other Instructors	
Year	Instructor*	Average	n	Average	n
2017	MI	4.67	90	4.58	90
2017	FI	4.44	90	4.58	90
2016	MI	4.23	86	4.13	86
2016	FI	4.08	86	3.67	86
2015	PMI-1	3.32	68	2.90	68
2014	PMI-1	3.00	53	2.40	53
2013	PMI-1	3.73	55	3.31	54
2012	PMI-1	3.34	50	3.06	50
2011	PMI-2	3.14	63	2.68	63
2010	PMI-2	3.62	53	3.30	53
2009	PMI-2	2.93	40	2.45	40

*MI = current male instructor, FI = current female instructor, PMI-# = past male instructor, different from the current MI

The following open-ended questions from the SETs were analyzed:

1. *"What did you especially like about the way this instructor taught the course?"*
2. *"What suggestions would you make to improve the way this instructor taught the course?"*
3. *"What did you especially like about this course?"*
4. *"What suggestions would you make to improve this course?"*
5. *"Please write any additional comments you may have below."*

The open-ended student responses were coded by following a qualitative coding analysis. Each student response was given a unique number to allow for a blind analysis. To prevent instructor identification and researcher bias, a third-party research assistant assigned single-blind identification numbers to each response and de-gendered the instructor of each response. Duplicates and nonsense answers were removed by the third party. If there were duplicate answers, they generally were the same response to a question for both instructors in the same year. Once this correction was done, roughly 10% of the remaining sample was initially analyzed to create a codebook for the responses for a semi-grounded approach.[35] The initial screening and generation of codes were based off Basow's work.[36] Seven different categories were generated, with positive and negative coding in each category. Those categories were Scholarship (including positive phrases such as "knowledgeable about the material," "great learning strategies," and "insightful answers" and negative phrases such as "you do have a PhD," "offer better help," and "too many mistakes"), Organization and Clarity (positive phrases such as "thorough on materials," "prepared," and "impressive time management" and negative phrases such as "disorganized," "not legible," and "needs more preparation"), Quality of Personality (positive phrases such as "short and to the point," "patient," and "respectable" and negative phrases such as "too apologetic," "annoying," and "condescending"), Instructor-Group (positive phrases such as "walked us through each step," "good at encouraging class," and "[positive] class involvement"

and negative phrases such as "it [interaction] forced us to rush," "more guidance...to team," and "[need] help with delegating in a group"), Instructor-Individual (positive phrases such as "successful in making me learn," "helped me see," and "genuinely care about the students" and negative phrases such as "[I was] left a bit confused" and "answers...confusing"), Dynamism/Enthusiasm (positive phrases such as "effort," "enthusiastic," and "cares a lot" and no negative phrases found), and Overall General Comments (either positive or negative). In addition, if any comments specifically pointed out shortcomings between the instructors, those comments were placed in a special category and analyzed.

Two other third-party analysts returned results after coding the entire set of student comments that remained after duplicates were removed, and performed an inter-rater reliability (IRR) test to confirm appropriate analyses.[37] Of the five open-ended questions, only questions 1 and 2 had similar numbers of unique responses. Questions 3 and 4 had many duplicate responses, signifying that students copied and pasted responses for both instructors into these questions. Question 5 had only a few duplicates and was considered in the coding analysis. As such, the IRR test was performed on Q1, 2, and 5 by randomly sampling 20 comments from each section and having both the researcher and two student analysts code them. The coding choices were compared, and the percentages for each category were reported. Figure 1 below illustrates the filtering of open-ended responses.

Both 2016 and 2017 responses were analyzed independently and then grouped together for analysis. The responses for each instructor were also compared between 2016 and 2017 to determine whether the pattern of comments for either instructor had changed from year to year.

The Classroom Observation Protocol for Undergraduate STEM (COPUS) was used to characterize the teaching practices of the instructors.[38] There were two different observation days for each faculty member to capture typical classroom days. The COPUS tool was used to record what the students and instructors were doing every two minutes during the 50-minute session. The data are presented in the Results section. The task character (usually incorporated into COPUS) was beyond the scope of this project, and so data were not collected in this area. The observation was conducted by a third party trained in COPUS.

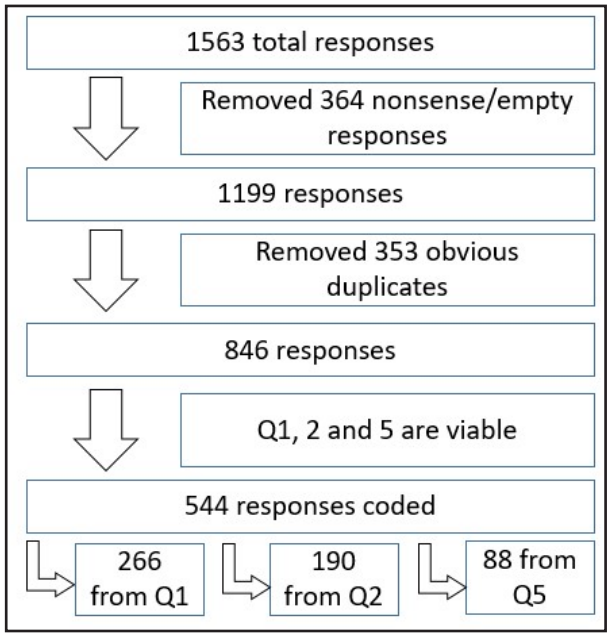

Figure 1. *Breakdown of Individual Response Filtering Process for both years.*

RESULTS

The following historical data shown in Table 1 from previous all-male instructors of the course were collected to provide a reference point between the instructors in the study and the previous instructors of the course. Note that the class size has grown over time, and all student responses were on a scale of 1 to 5 for all years.

The following results were obtained from the Likert data using Student's t-test. Table 2 presents the average scores for each of the Likert categories, given in mean ± standard deviation. Table 3 displays the calculated p-values from Student's t-test.

Tables 4 and 5 contain the summary reports from the coding analysis. The categories, as noted in the Methods section, were Scholarship, Organization and Clarity, Quality of Personality, Instructor-Group, Instructor-Individual,

TABLE 2
Means from categories, with 5.0 being the maximum possible score

	Instructor*	2016 (n=86)^	2017 (n=90)^	Combined (n=176)^
Overall Effectiveness	MI	4.23 ± 0.08	4.67 ± 0.06	4.45 ± 0.05
	FI	4.08 ± 0.08	4.44 ± 0.07	4.27 ± 0.06
Team Effectiveness	MI	3.86 ± 0.09	4.63 ± 0.06	4.26 ± 0.06
	FI	3.86 ± 0.09	4.63 ± 0.06	4.26 ± 0.06
Comparison to other instructors	MI	4.13 ± 0.09	4.58 ± 0.06	4.36 ± 0.06
	FI	3.67 ± 0.09	4.58 ± 0.06	4.14 ± 0.06
Comparison of Team	MI	3.71 ± 0.10	4.59 ± 0.06	4.16 ± 0.07
	FI	3.82 ± 0.10	4.59 ± 0.06	4.19 ± 0.07
Respect	MI	4.63 ± 0.06	4.86 ± 0.04	4.74 ± 0.04
	FI	4.63 ± 0.06	4.86 ± 0.04	4.74 ± 0.04

^Means are given in the format: mean ± standard deviation; *MI = male instructor, FI = female instructor

Dynamism/Enthusiasm and General Comment. Of the five questions, only Questions 1 and 2 had a sufficient number of viable responses for comparison of the male and female instructor comments. Many duplicate responses were given for Questions 3 through 5, where students copied the same response verbatim for the male and female instructors. As a result, questions 3 and 4 were not considered in the coding scheme; for question 5, only the 2016 data were coded because 2017 had a complete set of duplicates between instructors. Table 4 shows the total counts and proportions (in percentages) of positive and negative comments for the instructors.

The following treemaps were created to illustrate the coded data. A breakdown of all categories, along with the original comments, is provided in the supplemental documents. Figures 2 and 3 depict the numbers of coded responses in each category across both years for the male instructor and for the female instructor, respectively. The results are reported as the positive or negative counts for each category.

Table 5 lists the percentages from the inter-rater reliability test between the researcher and the third-party verification.

The COPUS data that were collected are shown in Figure 4. Four total class periods were used to collect the COPUS data, two for each instructor out of the 45 days of in-class time. A key to the COPUS data is provided in Table 6; note

TABLE 3
Significance of Likert Scores via P-values resulting from Student's t-tests

Question†	Overall Effectiveness	Team Effectiveness	Comparison to other instructors	Comparison of Team	Respect
2016 MI to 2016 FI	0.096	0.500	<0.001*,^	0.219	0.500
2017 MI to 2017 FI	0.008*,^	0.500	0.500	0.500	0.500
2016 MI to 2017 MI	<0.001*,^	<0.001*,^	<0.001*,^	<0.001*,^	0.001*,^
2016 FI to 2016 FI	0.001*,^	<0.001*,^	<0.001*,^	<0.001*,^	0.001*,^
Combined 2016 and 2017 MI to FI	0.007*,^	0.500	0.005	0.362	0.500

* α=0.05; ^Bonferroni adjusted α=0.01; †MI = male instructor, FI = female instructor

TABLE 4
Total counts and proportion of positive and negative comments across all categories

		Total Count		Proportions (%)	
Year	Instructor*	Positive	Negative	Positive	Negative
2016	MI	183	84	69	31
2016	FI	163	75	68	32
2017	MI	209	61	77	23
2017	FI	203	68	75	25
Overall	MI	392	145	73	27
	FI	366	143	72	28

*MI = Male instructor, FI = Female instructor

TABLE 5
Inter-rater reliability between the researcher and third-party student analysts, in %

Scholarship		Organization and Clarity		Quality of Personality		Instructor-Group	
Researcher-Student 1	Researcher-Student 2	Researcher-Student 1	Researcher-Student 2	Researcher-Student 1	Researcher-Student 2	Researcher-Student 1	Researcher-Student 2
82.5	80	80	77.5	92.5	90	97.5	90
Together:	81.25	Together:	78.75	Together:	91.25	Together:	93.75
Instructor-Individual		Enthusiasm		General			
Researcher-Student 1	Researcher-Student 2	Researcher-Student 1	Researcher-Student 2	Researcher-Student 1	Researcher-Student 2		
90	75	95	92.5	100	60		
Together:	82.5	Together:	93.75	Together:	80		

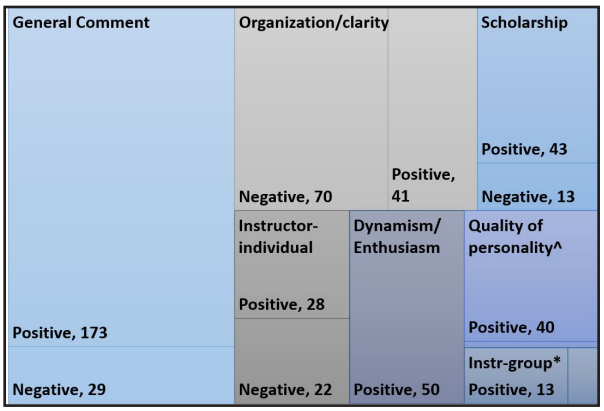

Figure 2. *Treemap for the male instructor across the coded categories. ^Quality of Personality: Negative = 2, *Instructor-Group: Negative = 4.*

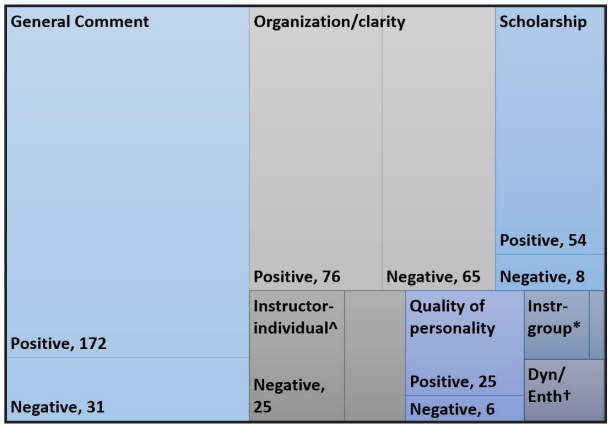

Figure 3. *Treemap for the female instructor across the coded categories. ^Instructor-Individual: Positive = 16, *Instructor-Group: Positive = 9, Negative = 2, †Dynamism/Enthusiasm: Positive = 10, Negative = 0.*

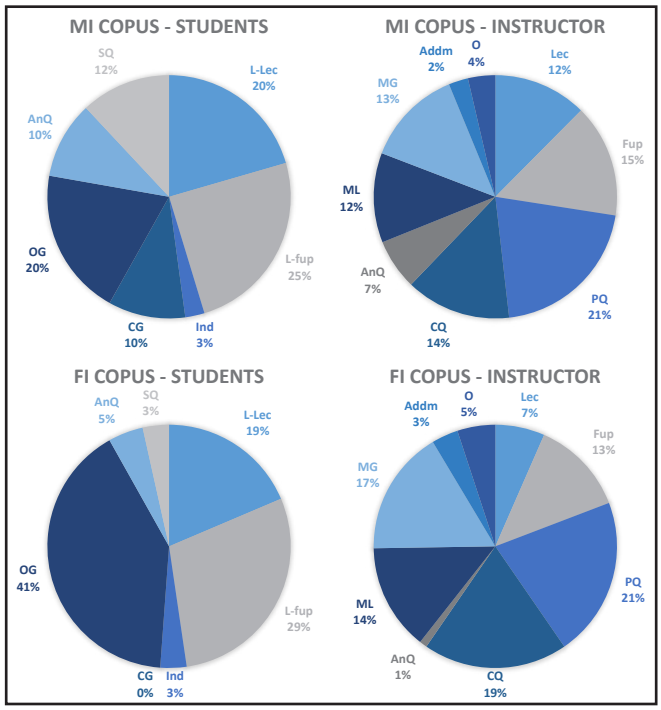

Figure 4. *COPUS results from both the male instructor and the female instructor on two separate days**
**MI = Male instructor, FI = Female instructor*

TABLE 6
COPUS Observation Codes

1. Students are Doing

L-Lec Listening to instructor's lecture.

L-Fup Listening to instructor's follow-up to small-group activity.

Ind Individual thinking/problem solving.

CG Discuss clicker question in groups.

OG Other assigned group activity.

AnQ Student answering an instructor's question to the whole class.

WC Engaged in whole-class discussion.

SQ Student asks question of the instructor.

2. Instructor is Doing

Lec Lecturing (presenting content, presenting a problem solution, etc).

Fup Lecture over content-related follow-up/feedback on CG to entire class.

PQ Posing non-clicker question to class. Assume rhetorical if no wait time or student response.

CQ Asking a clicker or polling question usually for polling of students

AnQ Listening to and answering student questions with entire class listening

ML Moving through class listening to ongoing student work during OG.

MG Moving through class guiding ongoing student work during OG.

Addm Administration, generally not content related

O Other – explain in comments.

that not all categories appear on the COPUS diagrams. Categories not present in Figure 4 had no data collected in them, meaning neither students or instructors performed those actions.

DISCUSSION AND CONCLUSION

Based on the Likert scale analysis of the SET scores and the evaluation of the open-ended comments, there seems to be a bias against female instructors in a male-dominant sophomore level chemical engineering course. When using SETs as a means of measurement, the instructors statistically differed in scores on overall effectiveness, indicating that females are slightly to moderately biased against.

Looking at the Likert comparisons shown in Table 1, in 2016, the male instructor was considered significantly better than other professors compared with the female instructor ($P<0.001$). In other words, the male instructor was seen as better than his peers more so than the female instructor was seen as better than her peers. Most of the students, at that point, had had no other female instructors in other core chemical engineering classes. Although they had not had any instruction from other female staff in the department, students could have interacted with female instructors from other departments, making it difficult to draw many conclusions regarding what genders the students may be using to compare the instructor.

The overall effectiveness in 2016 as shown in Table 2 was close to being significantly different, with a P-value of 0.096. The students rated the male and female instructors identically in terms of team effectiveness and respect toward students ($P=0.5$ and $P=0.5$). The components that each instructor contributed to the team respective to other academic teams showed that they viewed the male instructor as a better influence – but not significantly ($P=0.219$). Thus, this difference may be due to random chance. In 2017, the students rated the male instructor as having a significant positive difference in overall effectiveness ($P=0.008$). Thus, the students viewed the male and female instructors as significantly different; the male instructor was rated higher than the female instructor (4.67 ± 0.06 and 4.44 ± 0.07, respectively). In all other aspects, the two were rated identically, as most students reported the same Likert scores for both instructors ($P=0.5$) in these other categories. It should be noted that the male instructor had taught each student cohort in a core course the previous semester and so was already familiar to the students before the study course began. This confounding factor could not be designed out. Both instructors were highly rated compared to past instructors, and the measures taken to converge teaching methods attempted to minimize any effects of this. Future research should investigate these issues further.

When looking at the instructors from year to year, both

outperformed themselves from the previous year in all categories (P=0.001 or <0.001 in all categories), showing that they both were able to improve the students' perspectives of their teaching abilities from 2016 to 2017. When the 2016 and 2017 results were combined, there was a statistical difference between the students' view of the teaching effectiveness between the male and female instructors. The Likert analysis shows that for both years combined, the students viewed the teaching effectiveness of the two instructors as statistically different (P=0.007). The students also viewed the male instructor as much more effective compared with other instructors than the female instructor compared with other instructors (P=0.005). The difference in contribution to team effectiveness compared with other teams showed that the male instructor was rated higher, but not significantly (P=0.362). Both instructors were rated higher than the previous instructors historically, but no statistical tests were performed on these historical data. Because of the anonymity of the collected data, we are unable to associate the gender of the student with their ratings of the professors. As such, we can only take bulk parameters for the class as a whole, such as ratio of percent male to percent female, and draw general conclusions from the data.

In terms of the qualitative data, when we looked at both the male and female instructors' open-ended comments, there were similar patterns of positive and negative responses from the students. The proportion of comments that were coded as either positive or negative (or having a mixture of both) came out to be roughly equal for both instructors. An inter-rater reliability (IRR) test was used with two other coders to show that researcher bias was minimized. As seen from Table 6, there was considerable agreement between the researcher and the student analysts. The full coding analysis can be found in the supplemental material.[39] Between the male and female instructors, the male always had a slightly higher proportion of positive to negative comments overall (73% positive to 27% negative vs 72% positive to 28% negative, respectively). The biggest difference was primarily in the language used by the students. In particular, one student used profanity when creating an open-ended comment about the female instructor. In addition, when a student comment compared the two instructors, the female instructor was reviewed negatively for creating exams that were too difficult or grading more harshly than the male instructor. However, as stated in the Classroom Design section of the paper, the male instructor created the exams and the grading rubrics for evaluating those exams. The COPUS data provided in Figure 4 show the similarity between the two instructors' teaching practices and that the instructors did not differ significantly in teaching strategy, spending nearly the same amount of class time on each type of activity. Note that the COPUS data were collected over 2 class periods for each instructor out of the 45 total classroom events during the first year. The COPUS data gives insight into the teaching styles (and their similarities) but cannot be statistically distinguished as different with a sample size of two. With the collaboration, convergent teaching style, and controls placed on the study, it is difficult to justify that the variance in the SETs was from the difference in teaching instruction and experience between the male and female instructors.

Of our original hypotheses:

1. *The female instructor would be evaluated as a less effective instructor through Likert scale rankings of instruction and*
2. *The female instructor would receive more negative open-ended comments than the male instructor,*

only the first hypothesis was supported by the data collected for this study. The combination of the Likert data and the coded open-ended comments support the notion that female instructors experience at least slightly more negative bias than male instructors in male-dominated STEM fields such as chemical engineering. As this project is ongoing, we have replicated our work with another two instructors in two other classes, but we are not yet ready to analyze that data.

ACKNOWLEDGEMENTS

Thanks to Jonathan Cox for performing the COPUS analysis on both the male and female instructors during their class sessions.

REFERENCES

1. Goos M and Salomons A (2017) Measuring teaching quality in higher education: assessing selection bias in course evaluations. *Res. High. Educ.* 58(4):341–364.
2. Price L, Svensson I, Borell J and Richardson JTE (2017) The role of gender in students' ratings of teaching quality in computer science and environmental engineering. *IEEE Trans. Educ.* 60(4):281–287.
3. Anderson K and Miller ED (1997) Gender and student evaluations of teaching. *Polit. Sci. Polit.* 30:216–219.
4. Basow SA and Montgomery S (2005) Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *J. Pers. Eval. Educ.* 18(2):91–106.
5. Fenn AJ (2015) Student evaluation based indicators of teaching excellence from a highly selective liberal arts college. *Int. Rev. Econ. Educ.* 18:11–24.
6. Basow SA and Silberg NT (1987) Student evaluations of college professors: are female and male professors rated differently? *J. Educ. Psychol.* 79(3):308–314.
7. Wachtel HK (1998) Student evaluation of college teaching effectiveness: A brief review. *Assess. Eval. High. Educ.* 23(2):191–212.
8. Kaschak E (1978) Sex bias in student evaluations of college professors. *Psychol. Women. Q.* 2(3):235-243.
9. Leone-Perkins M, Schnuth R and Kantner T (1999) Preceptor-student interactions in an ambulatory clerkship: Gender differences in student evaluations of teaching. *Teach. Learn. Med.* 11(3):164–167.
10. Tieman CR and Rankin-Ullock B (1985) Student evaluations of teachers: an examination of the effect of sex and field of study. *Teach. Sociol.* 12(2):177–191.
11. Krautmann AC and Sander W (1999) Grades and student evaluations of teachers. *Econ. Educ. Rev.* 18(1):59–63.
12. Centra JA (2009) *Differences in Responses to the Student Instructional Report: Is It Bias?* (Educational Testing Service) Available at: https://

www.ets.org/Media/Products/SIR_II/pdf/11466_SIR_II_ResearchReport2.pdf.
13. Dennin M, et al. (2017) Aligning practice to policies: changing the culture to recognize and reward teaching at research universities. *CBE Life Sciences Education*: 16(4):es5. doi: 10.1187/cbe.17-02-0032
14. Stark P, Ottoboni K and Boring A (2016) Student evaluations of teaching (mostly) do not measure teaching effectiveness. *Sci. Res.*:1–11.
15. Nadler JT, Berry SA and Stockdale MS (2013) Familiarity and sex based stereotypes on instant impressions of male and female faculty. *Soc. Psychol. Educ.* 16(3):517–539.
16. Moore M (1997) Student resistance to course content: reactions to the gender of the messenger. *Teach. Soci.* 25(2):128–133.
17. Griffin BW (2001) Instructor reputation and student ratings of instruction. *Contemp. Educ. Psychol.* 26(4):534–552.
18. Spooren P, Brockx B and Mortelmans D (2013) *On the Validity of Student Evaluation of Teaching: The State of the Art* doi:10.3102/0034654313496870.
19. Miller J and Chamberlin M (2000) Women are teachers, men are professors: A study of student perceptions. *Teach. Sociol.* 28(4):283–298.
20. Anderson KJ and Smith G (2005) Students' preconceptions of professors: Benefits and barriers according to ethnicity and gender. *Hisp. J. Behav. Sci.* 27(2):184–201.
21. Bullard L and Felder R (2007) A student-centered approach to teaching material and energy balances. *Chem. Eng. Educ.* 41(3):167-176.
22. Walton GM, Logel C, Peach JM, Spencer SJ and Zanna MP (2015) Two brief interventions to mitigate a "chilly climate" transform women's experience, relationships, and achievement in engineering. *J. Educ. Psychol.* 107(2):468–485.
23. Prince M (2004) Does active learning work? A review of the research. *J. Eng. Educ.* 93(3):223–231.
24. Bell BS and Kozlowski S (2008) Active learning: Effects of core training design elements on self-regulatory processes, learning and adaptability. *J. Appl. Psychol.* 93(2):296–316.
25. Talanquer V, Bolger M and Tomanek D (2015) Exploring prospective teachers' assessment practices: Noticing and interpreting student understanding in the assessment of written work. *J, Res. Sci. Teach/* 52(5):585–609.
26. Rowden GV and Carlson RE (1996) Gender issues and students' perceptions of instructors' immediacy and evaluation of teaching and course. *Psychol. Rep.* 78:835–839.
27. Centra JA and Gaubatz NB (2000) Is there gender bias in student evaluations of teaching? *J. Higher. Educ.* 71(1):17–33.
28. Zimmerman DW (2004) Inflation of type I error rates by unequal variances associated with parametric, nonparametric, and rank-transformation tests. *Psicologica* 25:103– 133.
29. Zimmerman DW (2004) A note on preliminary tests of equality of variances. *Br. J. Math. Stat. Psychol.* 57(1):173–181.
30. Ruxton GD (2006) The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behav. Ecol.* 17(4):688–690.
31. de Winter JCF and Dodou D (2010) Five-point likert items : t test versus Mann-Whitney-Wilcoxon. *Pract. Assessment, Res. Eval.* 15(11):1–16.
32. Norman G (2010) Likert scales, levels of measurement and the "laws" of statistics. *Adv. Heal. Sci. Educ.* 15(5):625–632.
33. Armstrong RA (2014) When to use the Bonferroni correction. *Ophthalmic Physiol. Opt.* 34(5):502–508.
34. Garamszegi LZ (2006) Comparing effect sizes across variables: Generalization without the need for Bonferroni correction. *Behav. Ecol.* 17(4):682–687.
35. Walther J, Sochacka NW, Benson LC, Bumbaco AE, Kellam N, Pawley AL and Phillips CML (2017) Qualitative research quality: A collaborative inquiry across multiple methodological perspectives. *J. Eng. Educ.* 106(3):398–430.
36. Basow SA (2000) Best and worst professors: Gender patterns in students' choices. *Sex Roles* 43(5/6):407–417.
37. Baillie C and Douglas EP (2014) Confusions and conventions: Qualitative research in engineering education. *J. Eng. Educ.* 103(1):1–7.
38. Smith MK, Jones FHM, Gilbert SL and Wieman CE (2013) The classroom observation protocol for undergraduate stem (COPUS): A new instrument to characterize university STEM classroom practices. *CBE Life Sci. Educ.* 12(4):618–627.
39. Supplemental Information. http://bit.ly/TCEPaper ❐