An Analysis of Sex Differences in Computing Teaching Evaluations

Priscila Santiesteban
University of Michigan
Ann Arbor, Michigan, USA
pasanti@umich.edu

Madeline Endres
University of Michigan
Ann Arbor, Michigan, USA
endremad@umich.edu

Westley Weimer
University of Michigan
Ann Arbor, Michigan, USA
weimerw@umich.edu

ABSTRACT

Anonymous student teacher evaluations are commonly used to evaluate the quality of computing instructors at the university level. However, such teaching evaluations are subject to gender and sex-based biases, calling into question their utility and scope. In this paper, we first use data from a large public American university to replicate previous findings showing that significant sex-related differences persist in computing teaching evaluations. Intriguingly, we find that the sex-differences in computing teaching evaluations are primarily driven by bias involving professors, while significant sex-based differences for student-instructors are not observed. Finally, we place the magnitude of the sex-based differences we observe into a broader engineering context.

CCS CONCEPTS

• **Social and professional topics → Gender**; **Computing education**.

KEYWORDS

teaching evaluations, sex, computer science education

ACM Reference Format:
Priscila Santiesteban, Madeline Endres, and Westley Weimer. 2022. An Analysis of Sex Differences in Computing Teaching Evaluations. In *Third Workshop on Gender Equality, Diversity, and Inclusion in Software Engineering (GE@ICSE'22), May 20, 2022, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3524501.3527604

1 INTRODUCTION

Anonymous student evaluations of teachers are intended to fairly evaluate the teaching quality of faculty members and help improve their current teaching methods [16]. Furthermore, student teacher evaluations are often used as evidence in faculty tenure and promotion decisions [13]. However, extensive research on the role of gender and race on student feedback has uncovered that it is often biased against women and minority faculty [2, 3, 8, 12, 14, 15]. Research suggests that female instructors and professors are consistently rated lower than their male counterparts in the same field.

This sex-based bias in student teaching evaluations is also apparent in computer science and software engineering departments in particular. A study conducted by Gorden *et al.* on the role of race and gender in student evaluations submitted on RateMyProfessors[1] found that women are generally rated lower than men in overall teaching quality in CS [8]. Such gender-related biases may be of particularly detrimental effect in computing, as the quality of student teacher evaluations could be crucial for broadening participation in a field where women are historically underrepresented.

We aim to further examine the existence and magnitude of gender-biases in computer science student teacher evaluations by conducting a similar study to that of Gorden *et al.* within the University of Michigan's Computer Science and Engineering division. This institution is a large, American public university producing over 1,000 undergraduate computing majors per year. Michigan's status as a large public university makes it an ideal case study for observing such biases: as of 2019 in United States, public universities confer 88% of associate degrees and 67% of bachelor degrees.[2] Furthermore, unlike some of the previous work that focuses on reviews on external sites, we analyze internal metrics that are directly used in tenure and promotion cases. We note that while previous focused on gender biases, our study uses biological sex as a proxy for gender: the employment data available only includes sex.

Overall, we analyze around 9,000 computing-related student teacher evaluations submitted in the Fall term of 2019 at the University of Michigan. We first replicate previous work, finding that instructor sex significantly and substantially correlates with aggregate evaluation score: male computing instructors consistently receive higher student teacher evaluation scores than do female instructors with a small-to-medium sized effect ($p = 0.03$, $d = 0.45$). We hypothesize that these differences may be due to student perception of competence [5]. Students may, unaccountably, hold female professors to a higher standard as their professional status changes and thus, expose them to tougher judgement than males.

Intriguingly, however, we also find that sex-related differences are driven fully by professors: we do not observe any significant sex-related differences in the evaluation scores for student instructors (e.g., undergraduate or graduate teaching assistants). Finally, we compare computing instructor student evaluation scores to those of non-computing engineering instructors at the same institution. We find that, contrary to our findings for computing, we do not observe significant sex-related differences in teaching evaluation scores for non-computing engineering instructors. This suggests

GE@ICSE'22, May 20, 2022, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-9294-5/22/05…$15.00
https://doi.org/10.1145/3524501.3527604

[1] https://www.ratemyprofessors.com/: Rate My Professors is a online site where students can submit publicly-available anonymous reviews of their professors.

[2] https://nces.ed.gov/programs/coe/indicator/cts

that sex-biases in student teacher evaluations are an issue that is particularly prevalent in computer science. Overall, our results provide additional evidence of continuing sex-related bias in computing student teaching evaluations, biases that should be considered when evaluations are used for decisions for hiring, promotion, and tenure.

2 BACKGROUND

The relationship between gender biases and teaching evaluations has been long studied in academia. Awareness of the lack of female representation in STEM-related fields has spiked concerns about the existence of gender biases within STEM. However, while the issue is well-explored for STEM in general, fewer studies explore these discrepancies in computer science and engineering in particular.

Price *et al.* examined ratings that spanned over 10 academic years at a large Sweden university where they found professors in subjects not typical for their gender were rated lower than those in subjects typical for their gender [15]. Specifically, they noted these differences in two subjects: CS and electrical engineering, which are known to be mostly male-dominated fields. Gordan *et al.* examined gender and racial bias in CS by looking at over 39,000 CS professors and their respective students evaluations on RateMyProfessor, a public website used by students to rate professors from their institutions [8]. They found female CS professors were scored lower than their male counterparts. Their findings and methodology have influenced the study presented on this paper.

Several studies have examined teaching quality and its relationship to student teaching evaluations that included CS faculty. Felton *et al.* studied the relationship between the perceived easiness, sexiness and teaching quality of professors [6, 7]. Bangert used a sample of 809 undergraduate and graduate students to validate online teaching evaluations, and the sample included CS students [1]. Unlike this analysis, these studies did not primarily focus on CS faculty.

3 METHODOLOGY

Data set: To understand sex differences in computing teaching evaluations, we analyze teaching evaluations submitted from Fall 2019 for engineering instructors at the University of Michigan, a large American public university. Our data set consists of the official anonymous teaching evaluations for all engineering instructors, of which around 300 are professors while the remaining 500 are teaching assistants (both graduate and undergraduate).[3] Of these instructors, around a third teach courses in Electrical Engineering and Computer Science (EECS). We note that while we are primarily interested in gender differences, the data available from Michigan only report sex. We thus use sex as a imperfect proxy for gender in our analysis. Furthermore, as teaching evaluations at Michigan are generally optional, we restrict our analysis to those courses in the top quartile of enrollment (> 40) to ensure enough student-submitted evaluations to permit meaningful analysis. We note, however, that due to the large size of the EECS department, this enrollment limit still captures the vast majority of EECS lecture sections. Table 1 contains a sex breakdown of the number of instructors in this top quartile that we include in our analysis for both the engineering department as a whole and for EECS in specific.

All Engineering Instructors	Professors	Non-Professors	All
Female	47	37	84
Male	137	103	240
Total	184	140	327
EECS Instructors	Professors	Non-Professors	All
Female	11	22	33
Male	48	56	104
Total	59	78	137

Table 1: Breakdown by sex of numbers of overall engineering instructors and computing ("EECS") instructors in specific who taught a class with enrollment over 40. Numbers are further broken down by status as a professor (e.g., research professors or faculty lecturer) or non-professor (e.g., student teaching assistant or Ph.D. student primary instructor).

Evaluation score calculation: For each instructor, teaching evaluations contain Likert-style answers of student agreement or disagreement with various statements such as "overall, the instructor was an excellent teacher", and "I knew what was expected of me in this course". We include all evaluation questions that directly relate to students' experiences in the course (see Table 2 for a full list). To obtain overall teaching evaluation scores, individual Likert answers are first converted to a 5-point scale, where 1 indicates strong disagreement with the statement and 5 indicates strong agreement. These numerical scores are then aggregated to create a final score for each instructor using an interpolated median,[4] a metric that retains more information about the distribution of responses than a traditional median. Final evaluation scores range from 1 to 5, with 5 being the best possible score. We calculate overall instructor evaluation scores for both each individual evaluation question and for all questions in aggregate.

Statistical methods: For the majority of our statistical significance tests, we use the standard two-tailed Student's t-test. However, as the Student's t-test assumes equal variance (homoscedasticity), so we use a heteroscedastic variant of the t-test when the ratio between the two group's variances is greater than 3, an established best practice.[5] For effect-size, we use Cohen's d, and to correct for multiple comparisons, we use the Benjamini-Hochberg adjustment [4], a method that better accounts for correlated significance results than the conservative Bonferroni adjustment. All results reported as significant are also significant after correction for multiple comparisons unless noted otherwise.

4 RESULTS

We organize our analysis around the following questions:

- *RQ1—CS Sex Differences:* Is the sex of computer science instructors correlated with average teaching evaluation scores?

[3] Our de-identified dataset is available at https://github.com/CelloCorgi/GE_2022 with results aggregated by course level to further protect anonymity; UM IRB HUM00214059. We are particular grateful to Mandy Barnes (Registrar's Office) and Joy Ma (Human Resources Information Services) for assistance.

[4] http://aec.umich.edu/median.php

[5] https://data.library.virginia.edu/a-rule-of-thumb-for-unequal-variances/

- *RQ2—Instructor Type:* Do sex-based differences present differently for professors vs. student instructors?
- *RQ3—Engineering Comparison:* How do sex differences observed with computing instructors compare to those observed with non-computing engineering instructors?

4.1 RQ1—CS Evaluation Sex Differences

We find that there *are* statistically-significant differences between male and female computing instructors' teaching evaluation scores, both in aggregate and also for one third of individual questions ($p < 0.05$, $q = 0.1$). For all sex-based teaching evaluation differences that reach statistical significance, male instructors receive higher scores than female instructors. For example, male instructors' average aggregate teaching evaluation score was 4.34 compared to only 4.23 for female instructors ($p = 0.02$). While 0.17 may seem minor, it represents a small-to-medium effect (Cohen's $d = 0.45$) and impacts many instructors. Furthermore, we find that some individual evaluation questions had even larger sex-correlated differences. For example, students are much more likely to agree that a course advanced their understanding of the subject matter if it had a male instructor (medium-sized effect, $d = 0.59$). Additionally, we note that this finding is not an caused by disparate levels of teaching evaluation submission rates (teaching evaluations at Michigan are optional). We observe no significant differences in the teaching evaluation submission rates for male and female professors (46% vs. 52%, $p = 0.18$). Our full statistical results for RQ1 can be found in the *Overall* columns in Table 2.

Our results align with, and support, prior findings of gender-based differences and bias in computing teaching evaluations [8, 15]. For example, Gordan *et al.* found that women's RateMyProfessor evaluations were consistently lower in score and contained fewer positive personality attributes than men's, a gap that that the authors attribute to systemic bias against women [8]. However, as with the prior work [8, 15], the effect-size of the sex-based differences we observe are small (with a few notable exceptions, our Cohen's d is less than 0.5). Thus, while we provide additional evidence of sex-based bias against female computing instructors in evaluations, this effect likely does not fully explain (though it may contribute to) gender-based success differences in computing academia.

4.2 RQ2—Differences by Instructor Type

We also investigate whether an instructor is a professor or not impacts observed sex-related differences. We define *Professors* as instructors who are either tenure-track research professors or teaching-track lectures while we define *Non-Professors* as undergraduate and graduate student-instructors. This investigation is motivated by a desire to isolate the magnitude of sex-related biases faced by professors as this statistic is the most likely to directly influence promotion and tenure decisions [13]. Furthermore, we were interested in potentially observing the effect of recent efforts at the University of Michigan to improve the process of hiring student instructors (see Kamil *et al.* [9] for one example, focusing on demographically-balanced hiring).

The *Professors* and *Non-Professors* columns in Table 2 contain our full statistical results. Overall, we find that sex-related differences in teaching evaluation scores are statistically significant, while differences for non-professor student instructors are not. In particular, we observed sex-related differences with $p < 0.05$ for six out of eight individual evaluation questions (three of which remain significant after multiple comparison correction) as well as for aggregate evaluation scores. For all of these significant comparisons, male instructors received higher scores. For non-professor student instructors, however, we observe no significant differences. Bias is also clear in the magnitude of the differences: the aggregate scores for professors differ by 7.25% of the 1–5 scale (Cohen's $d = 0.82$), compared to only a 2.25% difference for non-professors.

We propose three hypotheses that could explain this difference between bias against female professors and bias against female student instructors. First, it is possible that sex-based biases increase with the perceived authority of the instructor. This hypothesis could explain why no significant difference is observed for student instructors as they are students themselves, and thus, in some sense, are peers of the students giving the evaluations. However, we believe this hypothesis is unlikely to fully explain the observed variance: studies in the literature generally *do* observe sex-based bias for student-instructor evaluations [10]. Alternatively, it is possible the discrepancy is an artifact of our methodology as we focus on classes with enrollment over 40: compared to professors, student instructors are more likely to teach discussion sections, which in turn, are more likely to have lower enrollment. While more research is needed, we also find this hypothesis unlikely: though prone to bias due to the small number of evaluations submitted per course, a preliminary analysis finds no significant sex differences for the aggregate evaluation score for student instructors of classes with enrollment less than or equal to 40 ($p = 0.15$).

A third potential hypothesis is that sex-based bias is still evident for Non-Professors, however female student instructors actually *outperform* their male counterparts. If true, this may be due in part to a particular Michigan CS hiring practice that focuses on instructor quality (assessed via video recordings of explanations, etc.) rather than GPA and results in gender-balanced teaching assistant ratios for several undergraduate computing courses [9]. We encourage future studies to see if this difference between Professor and Non-professor instructors is replicated at other universities.

4.3 RQ3—General Engineering Comparison

Lastly, to place our results in context, we compare our EECS findings to those for non-computing engineering instructors at the same institution. First, we find that computing instructor teaching evaluation scores are higher than those of non-computing instructors ($p = 0.05$). Computing instructors had an average aggregate score of 4.36 while non-computing engineering instructors averaged 4.31, a small but significant difference.

However, an analysis of sex differences paints a more nuanced picture. While we found significant and substantial sex-correlated differences in aggregate evaluation scores for computing instructors ($p = 0.02$, $d = 0.45$), we do *not* find significant aggregate sex differences for non-computing engineering instructors ($p = 0.17$, $d = 0.23$). This result indicates that sex biases may be particularly prevalent in, and of greater magnitude for, computer science than they are for other engineering disciplines. Further research

	Professors			Non-Professors			Overall		
Question	F	M	p-value	Female	Male	p-value	F	M	p-value
Overall, this was an excellent course	3.80	4.25	**<0.01**	3.99	4.13	0.35	3.93	4.19	**0.02**
Overall, the instructor was an excellent teacher	4.07	4.40	0.04	4.35	4.36	0.91	4.26	4.38	0.16
The instructor explained material clearly	4.04	4.39	0.03	4.42	4.39	0.79	4.29	4.39	0.28
The instructor treated students with respect	4.65	4.69	0.65	4.60	4.63	0.40	4.61	4.66	0.28
The instructor seemed well prepared for class meetings	4.47	4.50	0.40	4.46	4.43	0.74	4.48	4.59	0.65
This course advanced my understanding of the subject matter	4.23	4.53	**<0.01**	4.28	4.40	0.18	4.26	4.46	**<0.01**
My interest in the subject has increased because of this course	3.84	4.27	**<0.01**	3.96	4.10	0.35	3.92	4.18	**0.01**
I knew what was expected of me in this course	4.03	4.28	0.05	4.16	4.18	0.84	4.12	2.23	0.18
Aggregate Evaluation	4.19	4.46	**0.02**	4.27	4.35	0.44	4.24	4.40	**0.03**

Table 2: Breakdown of EECS teaching evaluations by sex and instructor type. All teaching evaluation scores are on a scale from 1 to 5. Cells highlighted in green in the F and M columns indicate that the given value was significantly higher than that for the other sex. Cells highlighted in purple in the p-value columns are statistical tests for which $p < 0.05$. p-values that are also bold are those that remain significant after correction for multiple comparisons.

is needed to determine if this disciplinary difference is caused by Michigan-specific departmental characteristics or is indicative of broader computer science culture. However, it does provide evidence that accounting for sex-related bias may be of particular import when considering computing promotion and tenure cases at the university or engineering level.

5 SUMMARY AND CONCLUSION

Universities commonly use anonymous student teacher evaluations to evaluate the quality of computing and software engineering instructors at the university level. However, a rich literature shows that these teaching evaluations are subject to gender and sex-related biases that call into question their utility and scope.

Our investigation was partially motivated by a desire to support women professors in promotion cases or in letters of recommendation. While "folk wisdom" of gender bias in evaluation abounds and many can point to "flip the names"-style studies, we desired a concrete number for the magnitude of the bias in a large, public computing setting. Notably, some administrators expressed reluctance to admit evidence involving fewer students (e.g., $n = 43$ in MacNell *et al.* [11]) or evidence not specific to computing. As a result, we see value in a published study that focuses on quantifying bias in computing evaluations in particular.

Overall, we found additional evidence that sex-biases exist in computer science student teacher evaluations. Male instructors have significantly higher aggregate evaluation scores than female professors ($p = 0.03$, $d = 0.45$). Surprisingly, we found that these differences are driven by professors rather than student instructors: while we observe substantial and significant sex-related differences for computing professors ($p = 0.02$, $d = 0.82$), we do not observe any significant sex-related differences for student instructors. While more research is needed to see if this discrepancy is apparent at other institutions, it may be that this difference results in part due to Michigan-specific hiring practices for student instructors [9]. Finally, we observe that the magnitude of sex-biases in teaching evaluations is specific to computer science: we do not observe significant sex-related differences in the scores of non-computing engineering professors. While further work is needed to see if this result generalizes, it may be that sex-biases in teaching evaluations are an issue of particular import to the computer science community that should be considered going forward.

REFERENCES

[1] Arthur W. Bangert. 2008. The Development and Validation of the Student Evaluation of Online Teaching Effectiveness. *Computers in the Schools* 25, 1-2 (2008).

[2] Shelia K. Bennet. 1982. Student perceptions of and expectations for male and female instructors. 74, 2 (1982), 170.

[3] John A. Centra and Noreen B. Gaubatz. 2000. Is There Gender Bias in Student Evaluations of Teaching? *The Journal of Higher Education* 71, 1 (2000), 17–33.

[4] Shi-Yi Chen, Zhe Feng, and Xiaolian Yi. 2017. A general introduction to adjustment for multiple comparisons. *Journal of thoracic disease* 9, 6 (2017), 1725.

[5] Eric Deemer, Donna Thomas, and Candi Hill. 2011. Measuring Students' Perceptions of Faculty Competence in Professional Psychology: Development of the Perceived Faculty Competence Inventory. *Training and Education in Professional Psychology* 5 (02 2011).

[6] James Felton, Peter T. Koper, John Mitchell, and Michael Stinson. 2008. Attractiveness, easiness and other issues: student evaluations of professors on Ratemyprofessors.com. *Assessment & Evaluation in Higher Education* 33, 1 (2008).

[7] James Felton, John Mitchell, and Michael Stinson. 2004. Web-based student evaluations of professors: the relations between perceived quality, easiness and sexiness. *Assessment & Evaluation in Higher Education* 29, 1 (2004), 91–108.

[8] Nikolas Gordon and Omar Alam. 2021. *The Role of Race and Gender in Teaching Evaluation of Computer Science Professors: A Large Scale Analysis on RateMyProfessor Data.* Association for Computing Machinery, New York, NY, USA, 980–986.

[9] Amir Kamil, James Juett, and Andrew DeOrio. 2019. Gender-balanced TAs from an unbalanced student body. In *Proceedings of the 50th ACM technical symposium on computer science education.* 300–306.

[10] E Khazan, J Borden, S Johnson, and L Greenhaw. 2019. Examining Gender Bias in Student Evaluations of Teaching for Graduate Teaching Assistants. *NACTA Journal* 64, 2 (2019), 422–427.

[11] Lillian MacNell, Adam Driscoll, and Andrea N. Hunt. 2015. What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. *Innovative Higher Education* 40 (2015), 291–303.

[12] Herbert W. Marsh and Lawrence A. Roche. 1997. Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American psychologist* 52, 11 (1997), 1187.

[13] L.W. McCallum. 1984. A meta-analysis of course evaluation data and its use in the tenure decision. *Research in Higher Education* 21 (1984), 150–158.

[14] Deborah J. Merritt. 2008. Bias, the brain, and student evaluations of teaching. *St. John's Law Review* 82 (2008), 235.

[15] Linda Price, Ingrid Svensson, Jonas Borell, and John T. E. Richardson. 2017. The Role of Gender in Students' Ratings of Teaching Quality in Computer Science and Environmental Engineering. *IEEE Trans. Education* 60, 4 (2017), 281–287.

[16] Walter Schilling, John Estell, and Frederick Berry. 2011. Practical Interpretation of Student Evaluations for Starting Professors. In *American Society for Engineering Education.*