The Role of Race and Gender in Teaching Evaluation of Computer Science Professors: A Large Scale Analysis on RateMyProfessor Data

Nikolas Gordon
Trent University
Peterborough, Canada
nikolasgordon@trentu.ca

Omar Alam
Trent University
Peterborough, Canada
omaralam@trentu.ca

ABSTRACT

Recently, Computer Science (CS) education has experienced a renewed interest, driven by the demand in the fast-changing job market. This renewed interest created an uptick of enrollment in computer science courses. Increased number of students search for information about CS courses and professors. Often times, students turn to a professor's profile on online sites, e.g. RateMyProfessor.com (RMP), to read feedback and assessments made by other students. Student Evaluations of Teaching (SETs), conducted online or on paper, are widely used to assess and improve the teaching quality of professors, and to provide critical assessment of the teaching material and content. This paper studies the role of race and gender of computer science professors on their teaching evaluation by analyzing the publicly available data of over 39,000 CS professors on RateMyProfessor. We found that women are generally rated lower then men in overall teaching quality. They are also perceived lower in personality-related student feedback ratings, i.e. they perceived less humorous, and less inspirational. We also found that Asian professors are perceived to be tough graders and lecture heavy. They are also perceived to be more difficult in general.

CCS CONCEPTS

• **Social and professional topics → Student assessment**; **Computer science education**; **Software engineering education**.

KEYWORDS

Student Evaluation of Teaching, Race, Gender, Computer Science Education, RateMyProfessor

ACM Reference Format:
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. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (SIGCSE '21), March 13–20, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3408877.3432369

SIGCSE '21, March 13–20, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8062-1/21/03…$15.00
https://doi.org/10.1145/3408877.3432369

1 INTRODUCTION

Nowadays, Computer Science (CS) programs experience an increase in enrolment as students rush into them in record numbers [34]. This renewed interest on CS education is motivated by the US employment projections [2], which suggest that about three quarters of new job opening will be in computing. Usually, students seek to know about the professors they are taking courses with. They either ask previous students of the professor or view the professor's rating on online sites, such as *RateMyProfessor.com* (RMP). Despite the concerns with their usefulness and validity [37], Student Evaluation of Teaching (SETs) are the most common method for teaching evaluations in North American Universities [30]. SETs are extensively used in faculty's promotion and tenure evaluations. Despite criticism about their validity [36], research suggests that eliminating SETs is not the solution, and recommendations are made to mitigate these biases and formulate better teaching evaluation criteria [18].

Student Evaluation of Teaching (SETs) have been an active area of research. Several studies have been conducted to evaluate the validity of SETs, to which there was no clear answer [36]. Studies also have been conducted to understand effectiveness [6, 41], which finds that SETs are not related to a student's learning. Policy makers pay attention to these research, and they recommend to implement balanced metric evaluation systems of multiple measures [12]. Other research studied the scaling system used in SETs [13], which recommends that faculty should be allowed to suggest their own questionnaires and scaling system for student answers. Nevertheless, despite valid criticisms, SETs are widely used in evaluating teachers' performance.

Despite that SETs are widely being used, they are subject to racial and gender-based biases. Bias against women and minority faculty in SETs is well documented in literature [5, 9, 23, 26]. Research suggests that female professors consistently receive lower teaching evaluations than their male counterparts, despite the fact that students' grades or learning outcome are not impacted by the gender of the instructor [25]. In our study, we also find women are rated lower than men in overall teaching quality. In addition, race is an important factor that influences SET ratings. Because of racial bias, minority and black professors assume that there will be students who question their competence because of their race [17]. In this paper, we analyze more than 39,000 CS professor profiles from RMP and answer two research questions.

RQ_1 How do the overall teaching evaluations vary across race and gender of computing professors?

We find that CS professors are subject to racial and gender-based biases when it comes to their overall ratings on RMP. We find that

women are rated less than men on RMP and minority professors are rated less than their white colleagues. The difference between race and gender groups can be as high as 8%.

RQ_2 How does student feedback vary across race and gender of computing professors?

We compared CS professors based on the feedback tags on RMP. We identified number of race and gender-based differences. We found that women are rated lower than men in personality-related feedback, e.g. they are perceived to be less respected and less inspirational. We also noticed that Asian professors are perceived to be tough graders.

The rest of the paper is organized as follows. Section 2 presents related work. Section 3 outlines our approach and the data collection process. Section 4 discusses our results and findings. Section 5 discusses threats to validity and Section 6 concludes this paper.

2 RELATED WORK

Computer science professors are evaluated in different ways [8]. Different questionnaires are designed by different institutions to understand the quality of teaching of CS professors. In general, computer science professors receive low ratings in their teaching evaluation compared to other disciplines [31]. There are not many studies that focuses teaching evaluations for computer science professors. Felton et al. [14, 15] studied the relationship between the perceived easiness, sexiness and teaching quality of professors, including computer science professors. Bangert [3] used a sample of 809 undergraduate and graduate students to validate online teaching evaluations, the sample included CS students. We did not identify any work that studied the role of race and gender on the evaluations of CS professors.

Several studies investigated the factors that influence SET ratings, including the gender of the instructor. Johnson et. al. [19] found that factors, such as, class size, course level, gender, and academic rank of the instructor impacted the SET scores of engineering courses. Miles and House [27] found that small elective classes are correlated with positive SET ratings, while large classes that are taught by female instructors are correlated with negative SET ratings. Martin [24] finds that the perception of women in leadership role impacts SET rating when a woman teaches large classes. In a large study over 19,000 student evaluations, Mitchell et. al. [28] concludes that female professors consistently receive lower teaching evaluations than their male counterparts. Mitchell and Martin [28] conclude that women are evaluated differently in at least two ways, intelligence and personality. We found similar results in our study, i.e. students rate personality of women instructor's lower than men. In online teaching evaluations, students tend to provide bad teaching evaluation when they believe that the instructor is a female [16, 22]. Clayson [11] finds that student of business classes think that they will learn more from older male instructors. This was true for both male and female students.

In a large study over 19,000 student evaluations, Basow et. al. finds that African American professors were rated higher on student-group interactions [4]. In another study, black men scored the lowest on 26 multidimensional items that they studied [35]. In a recent paper, Chavez and Mitchell [10] conduct a limited study on the effect of race and gender on teaching evaluations of political science professors. They find that on average non-white faculty score 0.1 less on 5.0 scale than their white colleagues.

Table 1: Summary of the data used in our study

Category	#	Category	#
Professors	39561	White Men	10199
Universities	2869	White Women	3684
Men	26413	Black Men	168
Women	10086	Black Women	61
White	15009	Asian Men	3247
Black	245	Asian Women	1191
Asian	4788	Hispanic Men	779
Hispanic	1097	Hispanic Women	234

RateMyProfessor (RMP) has been a subject of number of studies. RMP and teaching evaluations administered by universities are found to effect professors equally [7]. Struber et. al. [38] analyze data of 500 professors to study the impact of a professor's gender on their SET. Rosen [33] conducted a large scale analysis on RMP data and found that professors in sciences are rated lower than their colleagues in humanities and arts. Theyson [40] studied the effect of perceived physical attractiveness of instructors on their overall rating on RMP, by analyzing the "Hot" tag on the site. RMP has removed tag, thus, we did not include it in the list of tags that we studied in RQ2. Kindred and Mohammed [20] find that the RMP ratings match real life experiences of students. Overall, the biases in SETs discussed earlier are also evident on RMP ratings [21, 32]. Subtirelu [39] found that Asian instructors face disadvantages related to race and language.

Our study differs from above discussed studies in two main ways. (1) We exclusively focus on CS professors. Although some studies compared sciences vs humanities professors' overall ratings as discussed above, we did not identify a study that focused on the role of gender and race on CS instructors. (2) We conduct a comprehensive analysis of all the feedback tags used in RMP. The closest study that we identified that studied feedback tags across race and gender was by Reid [32]. However that study focused on ratings from 25 liberal arts colleges and did not conduct detail analysis of the tags as we do on this paper.

3 APPROACH

3.1 Step 1: Data Collection

Data for the study was collected from the RMP website based on professors who are in Computer Science or Software Engineering departments. In general, the software engineering department is managed by the faculty of engineering and computer science is managed by the faculty of science. We chose these two departments because they often teach similar courses, i.e. courses in programming, databases, software engineering, and software design and modeling. RMP only contains data of professors from US and Canada. From our initial RMP website search query, we found 57,298 professors who teach CS courses. We wrote a web scraper to download the data, and our dataset contains student ratings/feedback as recent as of July 2020. Information on RMP that we collected included name, gender, school (university), discipline, feedback tags, and professor

ratings. RMP contains a number of associated tags for professors, such as *Would take again*, *Gives good feedback* and overall approval rating. These tags consist of simple statements that provide a general idea of what students believe their professor exhibit in their courses. We go into detail of all the tags later in this section.

Out of 57,298 professors, we were able to identify the gender of 36469 professors (26,413 men and 10,086 women). In terms of race, we were able to identify the race of 21,189 professors (15,009 White, 245 Black, 4,788 Asian, 1,097 Hispanic) based on their last names. In total, 19,563 professors had both their gender and race assigned. Table 1 provides a summary of the data used in this study.

3.2 Step 2: Gender Assignment

We assigned a professor's gender by analyzing the students comments on his/her page. Following other studies on RMP, e.g. [29], we assigned the gender based on pronouns. We counted the pronouns "he", "him", "her", and "She" on the student comments. We used a simple approach that summed up each occurrence of the pronoun above and placed them in their respective gender category. We then simply compared the two sets and then assigned the professor gender based on which pronoun set used most in his/her profile. However, following this approach, we were not able to assign the gender for 3152 professors. This is because the comments on the professor's profile on RMP contained zero number of pronouns or the number of pronouns for both genders broke even.

3.3 Step 3: Race Assignment

We assigned the professor's race based on their last name, following a previous approach [29], which used US census-derived likelihoods of race through last name analysis. The census-derived likelihood of race was added to each professor granted their last name exists in the likelihood study. The included races are White, Black, Hispanic, and Asian (including Middle Eastern, South, South-east, and Eastern Asian). Murray [29] assigned race to a professor based on 70% threshold. For example, an individual is assigned the race White, if at least 70% of individuals with the same last name are assigned the race White. Otherwise it will be left blank. We followed the same approach.

3.4 Step 4: Feedback Collection

RMP allows students to use 20 different feedback tags that describe a professor. Although most of the tags are self-explanatory, some are not straightforward to understand. We did not find any official documentation on RMP website that explains the meaning of each tag. We manually read students' comments that accompany the tags to understand them. Here is a brief explanation for each tag:

Giving good feedback: Describes a professor who provides extensive and valuable feedback on assignments and tests.

Respected: A professor who is highly regarded and respected by students.

Lot's of homework: A professor who assigns a lot of homework in the course.

Accessible outside of class: A professor who is available for questions outside class, i.e. through mediums such as emails and office hours.

Table 2: Precision, Recall and F1 measure

	Precision	Recall	F1 measure
Women	0.983	1	0.991
Men	1	0.991	0.995
Black	0.963	0.946	0.954
White	0.942	0.522	0.672
Asian	1	0.569	0.726
Hispanic	0.94	0.94	0.94
Black Women	0.962	0.929	0.945
Black Men	0.964	0.964	0.964
White Women	0.944	0.459	0.61
White Men	0.942	0.542	0.68
Asian Women	1	0.812	0.89
Asian Men	1	0.426	0.597
Hispanic Women	1	0.957	0.978
Hispanic Men	0.892	0.926	0.909

Get ready to read: Encourages students to do a lot of readings, or assigns a lot of reading assignments.

Participation matters: A professor who cares about students' in-class participation. It is important that students get involved in lectures and share their opinions and ask/answer questions.

Skip class? You won't pass: Attendance is part of grading scheme or attending lectures is important for understanding the material.

Inspirational: Inspires and motivates students to inquire more about the subject at hand.

Graded by few things: Very few graded items on the syllabus/heavily weighted assignments.

Test Heavy: Syllabus contains a substantial number of tests.

Group Projects: A professor that assigns group projects in the course.

Clear grading criteria: Straight-forward grading rubric.

Hilarious: Amusing lectures/interactions with students.

Beware of pop quizzes: Professor prefers to give quizzes without prior announcement.

Amazing lectures: Lectures that students enjoy.

Lecture Heavy: Lot of lecture hours with an ample amount of content.

Caring: Interacts with students in such a way that provides them reassurance and is sensible to students' situations/learning.

Extra Credit: Professor provides additional optional tasks that count for credit.

So Many Papers: A lot of written reports is expected.

Tough Grader: Strict on the grading criteria for assessments.

For each professor, we collected how many times a tag was reported. Then we collected statistics on each tag per race and gender group.

3.5 Step 5: Manual Analysis

Since we used an automated approach for race and gender assignment, we wanted to manually validate the results. We took a sample of 380 professor and manually checked their race and gender. Following other related research [32], we searched for the professor's

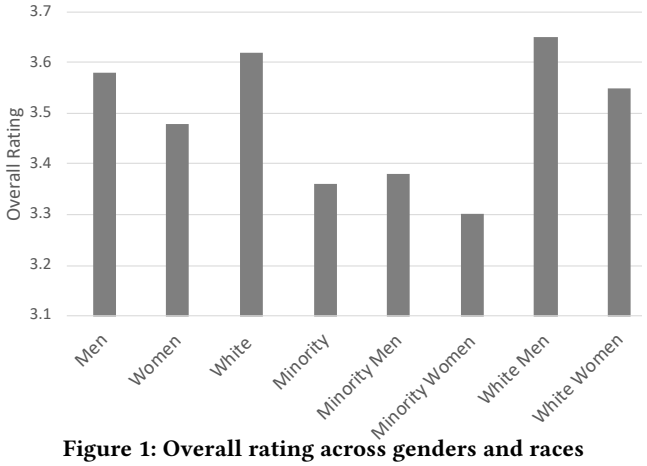

Figure 1: Overall rating across genders and races

Table 3: Overall Quality Ratings Across Race and Gender

Rating	Men	Women	Overall
Overall	3.57(1.58)	3.48(1.61)	3.55(1.59)
Black	3.44(1.64)	3.5(1.61)	3.46(1.63)
White	3.62(1.55)	3.54(1.58)	3.62(1.55)
Asian	3.31(1.61)	3.25(1.62)	3.31(1.61)
Hispanic	3.53(1.62)	3.44(1.65)	3.51(1.63)

Table 4: Difference between our findings and [32]

	White	Black	Asian	Hispanic
Women	-10%	-5%	-16%	-14%
Men	-9%	3%	-13%	-9%

profile online (i.e., on the university's website) and looked at their photos. We then calculated the precision and recall for each race/-gender group. Precision calculates the proportion of the identified positive class instances that truly belong to the positive class, i.e., Precision = $\frac{TruePositives}{TruePositives+FalsePositives}$. Recall calculates the proportion of identified positive class instances from all positive class instances in the dataset, i.e., Recall = $\frac{TruePositives}{TruePositives+FalseNegatives}$. F-measure is a metric that balances both precision and recall. Table 2 reports the precision, recall and F-measure for all gender and race groups. We notice that the precision of all groups is very high, which gives us the confidence on the automated approach used for gender and race assignment. However, we notice that recall suffers in number of occasions, e.g. for Asian and White professors. For example, the recall for White is 0.522. This is because there are many White professors who are assigned as *unknown* using the automated approach.

4 RESULTS AND DISCUSSION

4.1 RQ1: How do the overall teaching evaluations vary across race and gender of computing professors?

RMP provides an overall rating score for a professor. This rating is assigned by the students. We calculated the average overall scores

Table 5: Overall Difficulty Ratings Across Race and Gender

Rating	Men	Women	Overall
Overall	2.98(0.81)	2.83(0.80)	2.95(0.81)
Black	2.87(0.83)	2.93(0.82)	2.88(0.82)
White	2.96(0.80)	2.80(0.79)	2.91(0.80)
Asian	3.13(0.75)	3.03(0.74)	3.10(0.75)
Hispanic	2.90(0.87)	2.77(0.82)	2.85(0.85)

for different groups of gender and race: Men, Women, White, Minority (non-White), White Men, White Women, Minority Men, and Minority Women. In RMP, the overall quality is measured on a scale 1 to 5, with 5 being the highest quality. Fig. 1 shows the ratings for each group. As we can see, women are rated lower than men and minority professors are rated lower than their White colleagues. Within each White and Minority group, women are consistently rated lower than men. Percentage wise, women are rated 2% lower than men, while minority professors are rated 5.2% lower than their white colleagues. The difference slightly widens when we consider the gender and race of a professor, as minority women are rated 7% lower than white men.

Table 3 breaks down the minority group into different races. In particular, we have Black, Asian, and Hispanic as discussed before. For each gender/race group, we report the average overall quality rating and standard deviation in brackets. We did a series of ANOVA tests on different groups. The F-measure for gender was 27.34, and for race it was 43.61. When gender and race are used for groups, the F-measure was 18.69. All these values were at significance levels 0.01 and 0.05, and *p-value<0.00001*. The results of ANOVA tests suggest that the differences between means of the different groups are statistically significant. As shown in Table 3, when different races are considered, White professors are rated higher overall, followed by Hispanic professors, while Asian professors are rated the lowest. When both race and gender are considered, White men are rated the highest, followed by White women, while Asian women are rated the lowest. Among the minority groups, we noticed Hispanic men are rated the highest. The difference between White men and Asian women is 8%. We also notice that Black women are rated slightly higher than Black men. This is the only case where we found that women are rated higher.

Our analysis focused exclusively on computer science professors. We did not collect data of other fields. We wanted to compare whether the overall ratings of CS professors in this study are different from ratings from similar studies. We compared our findings with the findings by Reid [32], which analyzed the overall ratings of professors of liberal arts colleges. The study did not focus on a particular discipline and used RMP data for their analysis. Table 4 reports the percentage differences between the two studies. As we see on Table 4, computer science professors are consistently rated lower, except for Black men, where they are rated slightly higher. Asian computer science professors (both men and women) are rated considerably lower when compared with Asian professors in liberal arts colleges. This comparison confirms other studies that compared SET evaluations of science professors vs professors of other disciplines [31]. However, we notice that except for Black

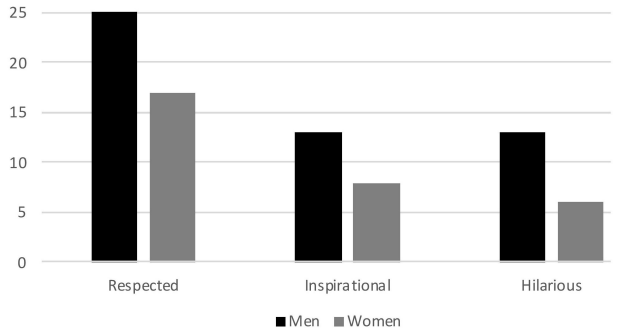

Figure 2: Gender Difference in Personality Tags

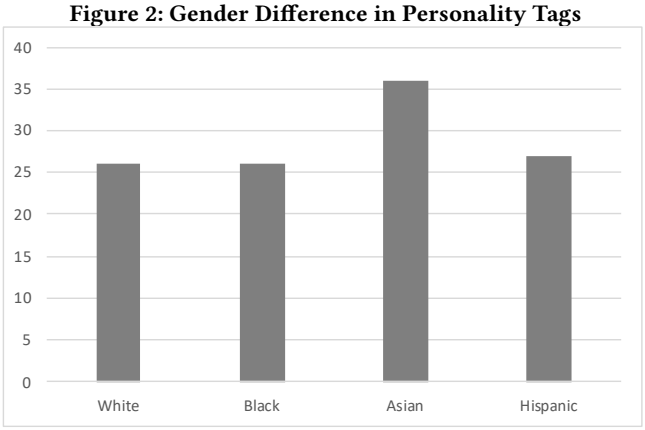

Figure 3: Tough Grader Tag Across Races

professors, CS professors of different race and gender groups are rated considerably lower.

Finally, RMP allows students to rate professors according to their difficulty, which measures the level of difficulty of a professor's lectures and assignments. Although, it seems that this metric is based on the content of material being taught, we noticed racial and gender based differences. Table 5 reports the difficulty level across gender and race groups. We notice that men are perceived more difficult than women. Among races, Asian professors are perceived to be most difficult.

Our analysis show that CS professors are subject to gender and race biases when it comes to overall quality ratings. This confirms the growing evidence from studies conducted on professors on different disciplines. In particular, we find that women are rated lower than men, and Asian professors are rated the lowest. The difference in overall rating between White men and Asian women is the highest. Among minority professors, Hispanic men fared better. We also find that CS professors in general are rated lower, when compared with similar studies conducted on professors from liberal arts colleges.

4.2 RQ2: How does student feedback vary across race and gender of computing professors?

In this question, we wanted to understand if race and gender of the instructor play a role on what kind of feedback he or she receives.

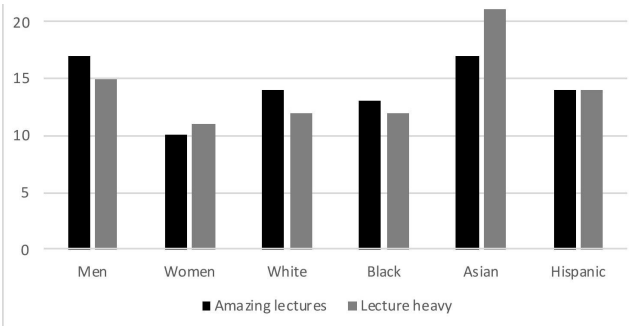

Figure 4: Lecture Tags Across Gender and Race

To answer this question, we calculated the number of occurrences of each feedback tag discussed in Section 3 across race and gender groups. Table 6 reports the averages of each group for tags. For example, for the tag *Gives good feedback*, 31% of men received this feedback tag from students. We also grouped related tags together as shown in the table. There are several observations that can be made from Table 6. We report them in the following:

Men are rated higher than women in Personality Traits. We noticed that personality tags, i.e. that describe the personality of a professor, are rated higher in men than in women. Fig 2 illustrates this observation. As shown in the figure, men are perceived as more respected, inspirational and hilarious. The differences between men and women are 8%, 5% and 7% respectively. When race is considered, Black and Asian women are perceived to be the least inspirational and hilarious. Only 4% of Black women are rated inspirational or hilarious. The percentages for Asian women were 8% and 4%, respectively.

Asian professors are perceived to be tough graders. Fig. 3 show that more than 35% of Asian professors are tagged with the *Tough Grader* tag. The percentage for other races are around 20%. In the Assessment group of tags, women are perceived slightly higher in three tags: *Extra credit, Clear grading criteria* and *Gives good feedback*. Among different races, Hispanics are rated higher in *Clear grading criteria* and *Gives good feedback*. These two tags are reported specifically higher for Hispanic women, more than 40% of them tagged with *Gives good feedback* and more than 35% are tagged with *Gives good feedback*.

Asian professors are perceived to give amazing lectures, but also they are perceived to be lecture heavy, as shown in Fig 4. White and Black professors have similar percentages for the lecture heavy tag. We also notice that men have higher percentages than women in of these two Lecture-related tags. When both gender and race is considered, we notice that White and Black women are reported to have least percentages for the *Amazing Lecture* tag.

Women are reported to be more caring than men. As shown in Table 6, 28% of women are tagged with the *Caring* tag, compared to 23% of men. Higher percentage of Hispanic women (27%) are tagged with the *Caring* tag, compared with women of other races (24%). In general, we did not notice gender difference when it comes to availability (*Accessible outside class* tag). However, when gender and race are both considered, we noticed that Black women are rated lower in availability, while Asian professors, both men and women, are rated higher.

Table 6: Student Feedback Tags Across Race and Gender Groups

Tag groups	Tags	Men	Women	White	Black	Asian	Hispanic	White men	White women	Black men	Black women	Asian men	Asian women	Hispanic men	Hispanic women
Assessment	Gives good feedback	31	32	31	29	28	33	32	33	31	28	29	30	33	42
	Test heavy	6	6	5	5	9	6	5	4	6	4	9	10	7	6
	Clear grading criteria	23	26	23	24	22	28	23	26	25	23	22	24	26	36
	Beware of pop quizzes	4	4	3	4	6	4	3	3	3	5	7	6	4	3
	EXTRA CREDIT	10	13	9	12	12	11	9	12	11	20	12	15	10	17
	Tough grader	29	28	26	26	36	27	27	26	28	25	38	37	29	28
Helping	Accessible outside class	13	13	13	12	15	13	13	13	14	9	15	17	13	13
	Caring	23	28	24	24	24	27	24	28	24	25	23	29	27	32
Homework and Tasks	Lots of homework	25	30	26	24	28	28	25	31	25	28	28	31	28	35
	Get ready to read	18	18	16	19	21	19	17	17	21	19	22	20	19	23
	Graded by few things	7	5	5	4	10	7	6	4	4	4	10	10	8	5
	Group projects	11	12	10	12	16	12	10	10	11	14	16	18	12	12
	So many papers	2	3	2	3	2	2	2	3	3	5	2	3	2	3
Lectures	Amazing lectures	17	10	14	13	17	14	17	9	15	9	19	13	16	11
	Lecture heavy	15	11	12	12	21	14	14	10	14	7	22	20	16	9
Participation	Participation matters	17	21	17	25	15	20	16	20	24	30	15	18	20	25
	Skip class? You won't pass.	23	21	20	20	25	23	21	20	23	15	26	26	24	24
Personality	Respected	25	17	22	22	22	23	25	17	24	17	25	18	26	21
	Inspirational	13	8	10	13	12	12	12	7	15	4	13	8	13	9
	Hilarious	13	6	12	11	9	9	15	6	14	4	10	4	11	6

Women are perceived to give a lot of homework. 30% of women have this tag compared with 25% of men. When race is considered, 35% of Hispanic women have this tag. Women are also rated higher in the group project tag. On the other hand, men are rated higher in the *Graded by few things* tag. Among races, Asian professors are rated the highest *Graded by few things*. They are also rated higher in the group projects tag. Asians, especially Asian men are also rated higher in *Get ready to read*.

For women, in class participation matters. 21% of women have this tag compared to 17% of men. Among races, higher percentage Black professors are reported to reward for in-class participation (with Black women having higher percentage of 30%), while Asian professors have the lowest percentage for this tag. On the other hand, men have higher percentage of *Skip class? You won't pass* tag, suggesting that actually attending the class matter more to them. Among races, Asian professors have higher percentage for this tag.

Gender and race of the professor play a role on how students provide feedback about them. When we analyzed RMP tags, we noticed that women are rated lower in personality-related tags, i.e. they are perceived to have less humor, to be less inspirational and to be less respected. We also notice that Asian professors are perceived to be tough grader, but they are also perceived to have amazing lectures and to be lecture heavy. In terms of homework, higher number of women are reported to give a lot of homework.

5 LIMITATIONS AND THREATS TO VALIDITY

• **Internal Validity** Our dataset contains a very low number of Black professors (168 men and 61 women). The lack of Black representation in computer science is well documented. According to a recent Computing Research Association survey, there were only 16 Black PhD graduates in the US in 2019 [1], which may explain the low number of Black professors in RMP.

We compared our findings in RQ1 with findings by Reid [32], which was conducted in 2010. Reid investigated the overall ratings of professors from 25 liberal arts colleges in the US. Although his dataset did not exclude any discipline, it is 10 years old and did not study professors from large number of universities as we did in this study. Further research is warranted to compare CS professors with professors of other disciplines.

• **External Validity** Threats related to generalization of results. We conducted our study on unofficial online student evaluations. It remains to be seen if our findings can be generalized to official teaching evaluations conducted by computer science departments. Furthermore, RMP only contains ratings of professors in US and Canada. Therefore, It remains to be seen if our findings can be generalized to other parts of the world.

• **Construct Validity** We used automated approach for gender and race assignments, which may lead to false positives and pose a threat to validity for our findings. We mitigated this concern by manually analyzing a sample of 380 professors. As discussed in Section 3, we received high precision for all gender/race groups.

6 CONCLUSION

This paper analyzed CS professors' ratings on *RateMyProfessor.com* to investigate whether the professor's race and gender play any role on their teaching evaluation. We collected data from over 39,000 professors and analyzed their overall ratings and feedback tags assigned by students. Our analysis reveals clear bias against women and minority professors. Women of all races are rated lower than men in overall teaching quality. They are also perceived to be less than men in personality-related student feedback. On other side, women are perceived to be more caring than men. We also noticed that professors from racial minority backgrounds receive lower overall ratings than their White colleagues. These findings adds to the growing evidence about race and gender-based biases on SETs. In future, we plan to compare our study with official SETs collected from computer science departments. We also plan to conduct in-depth analysis on students' written comments on RateMyProfessor. Finally, we plan to compare our findings with recent RMP ratings of professors across disciplines.

REFERENCES

[1] 2020. *CRA Survey*. https://cra.org/wp-content/uploads/2020/05/2019-Taulbee-Survey.pdf.

[2] Joel Adams. November 3, 2014. *Computing Is The Safe STEM Career Choice Today*. https://cacm.acm.org/blogs/blog-cacm/180053-computing-is-the-safe-stem-career-choice-today/fulltext

[3] Arthur W Bangert. 2008. The development and validation of the student evaluation of online teaching effectiveness. *Computers in the Schools* 25, 1-2 (2008), 25–47.

[4] Susan Basow, Stephanie Codos, and Julie Martin. 2013. The effects of professors' race and gender on student evaluations and performance. *College Student Journal* 47, 2 (2013), 352–363.

[5] Sheila K Bennett. 1982. Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology* 74, 2 (1982), 170.

[6] Anne Boring, Kellie Ottoboni, and Philip Stark. 2016. Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research* (2016).

[7] Stefanie S Boswell. 2016. Ratemyprofessors is hogwash (but I care): Effects of Ratemyprofessors and university-administered teaching evaluations on professors. *Computers in Human Behavior* 56 (2016), 155–162.

[8] Angela Carbone and Jens J Kaasbøll. 1998. A survey of methods used to evaluate computer science teaching. In *Proceedings of the 6th annual conference on the teaching of computing and the 3rd annual conference on Integrating technology into computer science education: Changing the delivery of computer science education*. 41–45.

[9] 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.

[10] Kerry Chávez and Kristina MW Mitchell. 2020. Exploring Bias in Student Evaluations: Gender, Race, and Ethnicity. *PS: Political Science & Politics* 53, 2 (2020), 270–274.

[11] Dennis E Clayson. 2020. Student perception of instructors: the effect of age, gender and political leaning. *Assessment & Evaluation in Higher Education* 45, 4 (2020), 607–616.

[12] Kevin Close, Audrey Amrein-Beardsley, and Clarin Collins. 2018. State-Level Assessments and Teacher Evaluation Systems after the Passage of the Every Student Succeeds Act: Some Steps in the Right Direction. *National Education Policy Center* (2018).

[13] Nida Denson, Thomas Loveday, and Helen Dalton. 2010. Student evaluation of courses: what predicts satisfaction? *Higher Education Research & Development* 29, 4 (2010), 339–356.

[14] 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), 45–61.

[15] 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.

[16] Alexandra N Fisher, Danu Anthony Stinson, and Anastasija Kalajdzic. 2019. Unpacking backlash: Individual and contextual moderators of bias against female professors. *Basic and Applied Social Psychology* 41, 5 (2019), 305–325.

[17] Roxanna Harlow. 2003. " Race doesn't matter, but...": The effect of race on professors' experiences and emotion management in the undergraduate college classroom. *Social psychology quarterly* (2003), 348–363.

[18] Therese A Huston. 2005. Race and gender bias in higher education: Could faculty course evaluations impede further progress toward parity. *Seattle J. Soc. Just.* 4 (2005), 591.

[19] Michael D Johnson, Arunachalam Narayanan, and William J Sawaya. 2013. Effects of course and instructor characteristics on student evaluation of teaching across a college of engineering. *Journal of Engineering Education* 102, 2 (2013), 289–318.

[20] Jeannette Kindred and Shaheed N Mohammed. 2005. "He will crush you like an academic ninja!": Exploring teacher ratings on ratemyprofessors. com. *Journal of Computer-Mediated Communication* 10, 3 (2005), JCMC10314.

[21] Angela M Legg and Janie H Wilson. 2012. RateMyProfessors. com offers biased evaluations. *Assessment & Evaluation in Higher Education* 37, 1 (2012), 89–97.

[22] 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, 4 (2015), 291–303.

[23] 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.

[24] Lisa L Martin. 2016. Gender, teaching evaluations, and professional success in political science. *PS: Political Science & Politics* 49, 2 (2016), 313–319.

[25] Friederike Mengel, Jan Sauermann, and Ulf Zölitz. 2019. Gender bias in teaching evaluations. *Journal of the European Economic Association* 17, 2 (2019), 535–566.

[26] Deborah J Merritt. 2008. Bias, the brain, and student evaluations of teaching. . *John's L. Rev.* 82 (2008), 235.

[27] Patti Miles and Deanna House. 2015. The Tail Wagging the Dog; An Overdue Examination of Student Teaching Evaluations. *International Journal of Higher Education* 4, 2 (2015), 116–126.

[28] Kristina MW Mitchell and Jonathan Martin. 2018. Gender bias in student evaluations. *PS: Political Science & Politics* 51, 3 (2018), 648–652.

[29] Dakota Murray, Clara Boothby, Huimeng Zhao, Vanessa Minik, Nicolas Bérubé, Vincent Larivière, and Cassidy R. Sugimoto. 2020. Exploring the personal and professional factors associated with student evaluations of tenure-track faculty. *PLOS ONE* 15, 6 (06 2020), 1–21. https://doi.org/10.1371/journal.pone.0233515

[30] Harry G Murray. 1984. The impact of formative and summative evaluation of teaching in North American universities. *Assessment and evaluation in Higher Education* 9, 2 (1984), 117–132.

[31] James S Pounder. 2007. Is student evaluation of teaching worthwhile? *Quality Assurance in Education* (2007).

[32] Landon D Reid. 2010. The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors. Com. *Journal of Diversity in higher Education* 3, 3 (2010), 137.

[33] Andrew S Rosen. 2018. Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of RateMyProfessors. com data. *Assessment & Evaluation in Higher Education* 43, 1 (2018), 31–44.

[34] Natasha Singer. January 24, 2019. *The Hard Part of Computer Science? Getting Into Class.* https://www.nytimes.com/2019/01/24/technology/computer-science-courses-college.html

[35] Bettye P Smith. 2009. Student Ratings of Teaching Effectiveness for Faculty Groups Based on Race and Gender. *Education* 129, 4 (2009).

[36] Pieter Spooren, Bert Brockx, and Dimitri Mortelmans. 2013. On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research* 83, 4 (2013), 598–642.

[37] Erin M Steffes and Lawrence E Burgee. 2009. Social ties and online word of mouth. *Internet research* (2009).

[38] Jenny M Stuber, Amanda Watson, Adam Carle, and Kristin Staggs. 2009. Gender expectations and on-line evaluations of teaching: Evidence from RateMyProfessors. com. *Teaching in Higher Education* 14, 4 (2009), 387–399.

[39] Nicholas Close Subtirelu. 2015. "She does have an accent but…": Race and language ideology in students' evaluations of mathematics instructors on RateMyProfessors. com. *Language in Society* 44, 1 (2015), 35–62.

[40] Katherine C Theyson. 2015. Hot or Not: The Role of Instructor Quality and Gender on the Formation of Positive Illusions Among Students using RateMyProfessors com. *Practical Assessment, Research, and Evaluation* 20, 1 (2015), 4.

[41] Bob Uttl, Carmela A White, and Daniela Wong Gonzalez. 2017. Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation* 54 (2017), 22–42.