Quantifying the Magnitude of Potential Bias in the Written Comments Presented in Student Evaluations of Teaching

Benjamin Ellis[1], Yunzhe Li[2], Dylan C. Friel[3], Daniel R. Jeske[4], Herbert K.H. Lee[5], Philip H. Kass[6]

Abstract

While bias in numerical scores of student evaluations of teaching is well documented, there has been less attention given to the potential bias in written comments corresponding to open-ended questions in the evaluations. We examine written comments from students at the University of California – Riverside and the University of California – Santa Cruz, analyzing them by gender and ethnicity and controlling for tenure status and the type of course (STEM versus not STEM). Our results demonstrate that there are combinations of the factors that show advantages for receiving a higher proportion of positive comments, but no evidence showing that the advantages skew in a consistent direction with respect to gender or race alone. In addition, the two campuses differ with respect to the combinations of factors that show advantages, with notable differences for male instructors of color at Santa Cruz. Considering the large literature on bias in numerical scores of student evaluations of teaching, our findings suggest that the written comments may be more appropriate for use in the evaluations of instructors than quantitative questions.

Key Words: student evaluations of teaching, gender and ethnicity bias, multinomial logistic regression, cluster analysis

[1] University of California, Riverside, belli004@ucr.edu
[2] University of California, Santa Cruz, yli566@ucsc.edu
[3] University of California, Riverside, dfrie001@ucr.edu
[4] University of California, Riverside, daniel.jeske@ucr.edu, corresponding author
[5] University of California, Santa Cruz, herbie@ucsc.edu
[6] University of California, Davis, phkass@ucdavis.edu

1. Introduction

Investigating potential for bias against instructors who are female or persons of color (POC) has been extensively studied in the literature for instructional evaluation questions with numerical scales (e.g., Ceci at al., 2023). There has been some evidence of bias in open-ended questions as well (Chávez and Mitchell, 2020), but there are relatively few studies of open-ended questions. In this study, we critically examine a large number of responses to open-ended questions on our campuses. We classified responses as clearly positive, clearly negative, or other (which include both mixed responses and those that do not directly comment on the instructor's teaching), and then examined correlations between these labels and demographics of the instructor.

Studying the potential for biases in student teaching evaluations is important because they are typically components of academic personnel reviews for faculty. If these evaluations are biased, then that should be considered in personnel reviews. Quantification of potential biases can provide context for how an instructor interprets their student feedback.

This paper is organized as follows. Section 2 provides a literature review and theoretical grounding for this work. Section 3 describes the methodology used for compiling the dataset of written comments and how they were labeled as positive, negative, or other. Section 4 presents the results of the correlation analysis that is based upon multinomial logistic regression analysis and cluster analysis. Our findings are presented in Section 5, and the paper concludes with a summary in Section 6.

2. Literature Review

At this point, there is a large body of work exploring biases in student evaluations of teaching. Ceci et al. (2023), Kreizer and Sweet-Cushman (2022), and Stoesz et al. (2022) each review a large number of such papers. A few specific examples are included here. MacNell et al. (2015) conducted an early experiment with four sections of an asynchronous online course where students did not actually see their instructor, and in some cases were told that they had a different instructor; students gave higher average ratings when advised their instructor was male, regardless of who the actual instructor was. Wagner et al. (2016) found significant bias against female instructors when accounting for a variety of covariates. Mangel et al. (2017) reported on a quasi-experimental setting of students randomly assigned to courses with male or female instructors, and noted that while student performance was comparable, student evaluations were lower on average for female instructors.

Biases may arise because of gender expectations held by students, where female instructors are penalized for not conforming to expected gender stereotypes, and while male instructors are not held to the same standards (Boring, 2016; Wagner et al., 2016; El-Alayli et al., 2018). In particular, women are expected to be more warm and caring, and students penalize female instructors who do not meet these expectations, while not holding these expectations for male instructors (Adams et al., 2022; Gelber et al., 2022).

Studies have also found biases against non-White instructors. Examples include Reid (2010) and Chávez and Mitchell (2020). Littleford et al. (2010) found that students perceive African American instructors as being more judgmental, which can impact their evaluations. Fan et al. (2019) demonstrate bias against instructors who are not native English speakers. Having a foreign accent can particularly drive bias (Subtirelu, 2015; Wang and Gonzalez, 2020). Fan et al. (2019) also finds that bias is less when a gender or ethnicity is better represented in a discipline.

Student evaluations have also been found to vary across subject areas, particularly in terms of STEM fields vs. non-STEM fields (Basow and Montgomery, 2005; Mengel et al., 2018). Thus it is important to consider controlling for discipline when analyzing evaluations, which we will do in our analysis.

Chávez and Mitchell (2020) found potential bias in open-ended questions, but the sample sizes were very small. Schmidt (2015) created a visualization tool for comments from the website Rate My Professor and demonstrated consistent differences in the positive and negative comments that male and female faculty received. Another line of research found that students have different expectations of male instructors and female instructors that appear in open-ended question responses, with potentially biased results for instructor evaluations (Mitchell and Martin, 2018; Adams et al., 2022, Gelber et al., 2022). Lindahl and Unger (2010) studied the cruel comments made toward instructors and found that these can be gendered, disadvantaging female instructors. Wallace et al. (2019) discuss intersectionality and the factors that may lead to bias against women of color.

It is worth noting that student evaluations have not been found to correlate with teaching effectiveness or student learning, and thus bias plays a larger role than teaching effectiveness (Carrell and West, 2010; Boring et al., 2016; Uttl et al., 2017).

3. Methodology

On each campus, access was obtained to student teaching evaluations. A set of courses that jointly existed at the University of California – Riverside (UCR) and the University of California – Santa Cruz (UCSC) was selected and written comments were extracted from the evaluations. Instructor names were redacted from individual comments via pattern matching. The campus faculty lead worked with a graduate student researcher on the data extraction. Undergraduate students were hired and trained to read through each of the responses to classify them as positive, negative, mixed, or other. Classification focused on the role and actions of the instructor. Each campus' IRB determined the study was exempt from formal review because it used existing anonymized institutional data.

Positive comments are those that are entirely positive, nearly all positive, or positive with some constructive feedback. Examples include "It was super helpful that <instructor-first-name> provided hand written notes as we go along during lecture," "The instructor is very clear on her expectations which allows the student to know what to prepare for" "Dr. <instructor-last-name> is always very enthusiastic and her lectures are usually very clear and easy to follow, and, although some of her spelling and information was mixed up sometimes, she will correct herself when made aware of it," and "Including drawing out the processes in some lectures was very helpful to me. The videos were also helpful although I wish those were made available after class in a link."

Negative comments are those that are critical or focus on areas of improvement. Examples include "It was hard to understand the testing style of the course," "give more practice. using the homework platform for everything is not useful," and "Exams were all free response but there was little to help us prepare for such difficult exams. Detailed study guides would have been immensely helpful."

Mixed comments contained both positive and negative elements, without being overwhelmingly positive or negative. Examples include "Study guide questions were greatly appreciated. Lectures weren’t as engaging and felt overwhelming,” “Lectures were full of information, but there was just a lot to cover, and it felt like it was covered rapidly.” “Videos are very helpful. I don't do the book readings because it is way too much reading,” and “Discussion sections assignments were somewhat engaging, but I didn't feel like it really amounted to substantive learning.”

Some comments were not directly related to the instructor’s teaching practices or course design choices, such as “I probably would have liked this class a lot more if it wasn't at 8 am.” “Nothing else to add,” “Lectures were very dense with information,” and “Merry Christmas.” These types of comments were merged with the mixed comments to form the “other” classification used in our analysis. A rationale for this is that we are primarily interested in the fractions of positive comments and negative comments as indicators of good and poor teaching. Also, the statistical model we use is more able to predict positive and negative outcomes, and less able to distinguish between mixed and indeterminate outcomes.

Each comment was read by two trained undergraduate students. If their labels agreed, then that was the final label. If they disagreed, then a third student would read and classify it, and a majority vote was used for the final label. In the rare instance that the three students gave three different labels, then the graduate student decided the final label, sometimes in discussion with the faculty member. The data we analyzed comprise a total of 22,319 comments at UCR and 85,833 comments at UCSC.

Occasionally we came across a comment that had strong potential to be sarcasm. These would get flagged for discussion with the graduate student and the faculty member, and if the meaning was unclear, these comments would be classified as “other.” There were very few of these, but one example is “Telling us to read the textbook was the most helpful piece of advice.” There is not enough information to know whether the student claimed it was a good textbook, or whether the instructor was perceived as so poor that the student felt that they learned everything from the textbook.

Our analysis focused on dependent variables that correspond to the percent of comments that are positive, negative, and other. We considered four covariates: instructor gender identity, instructor ethnicity, whether the instructor has tenure, and whether the class is a STEM class.

Because of the very small number of instructors who did not identify as either male or female, we only use those two categories. For ethnicity, we aggregate into White and POC categories, as disaggregating further can lead to very small numbers and potentially compromise the anonymity of instructors. Other research has found that evaluations in STEM fields are on average lower than those in non-STEM fields (Basow and Montgomery, 2005), so we controlled for that factor when evaluating potential bias by gender or ethnicity. The statistical analysis used multinomial logistic regression to assess the significance ($P < 0.05$) of the four covariates on the probability a comment is labeled positive, negative, or other. A subsequent clustering analysis was used to quantify the practical differences of the fitted trinomial distributions.

4. Statistical Analysis

4.1 Descriptive

In order to draw comparisons between the two campuses, we keep their data separate. Table 1 shows the distribution of the labeled comments from the UCR campus by gender (male versus female), ethnicity (POC versus white), tenure status (tenured versus not tenured), and discipline (STEM versus not STEM). Relative percentages for each of the 16 combinations of gender, ethnicity, tenure status, and discipline area correspond to an estimate of the trinomial distribution for the comment label outcome. Table 2 is the corresponding contingency table for the UCSC campus. The yellow shading in these tables will be described below.

			Not STEM					STEM			
			POC		White			POC		White	
			n	%	n	%		n	%	n	%
Positive	Not Tenured	Male	939	84.22%	872	67.7%	Male	555	66.95%	1067	54.58%
Other			105	9.41%	241	18.71%		180	21.71%	556	28.44%
Negative			71	6.37%	175	13.59%		94	11.34%	332	16.98%
Positive		Female	209	72.32%	185	98.4%	Female	219	78.49%	360	83.72%
Other			45	15.57%	2	1.06%		41	14.7%	57	13.26%
Negative			35	12.22%	1	0.53%		19	6.81%	13	3.02%
			POC		White			POC		White	
Positive	Tenured	Male	1323	86.58%	2157	89.88%	Male	1822	57.15%	2191	58.96%
Other			143	9.36%	160	6.67%		736	23.09%	817	21.99%
Negative			62	4.06%	83	3.45%		630	19.76%	708	19.05%
Positive		Female	834	80.97%	538	67.33%	Female	274	63.57%	573	62.76%
Other			97	9.42%	179	22.4%		66	15.31%	210	23.00%
Negative			99	9.61%	82	10.26%		91	21.11%	130	14.24%

Table 1. Classification Table of UCR Labels

			Not STEM					Not STEM			
			POC		White			POC		White	
			n	%	n	%		n	%	n	%
Positive	Not Tenured	Male	1024	58.35%	5879	70.31%	Male	2127	46.13%	5017	58.34%
Other			460	26.21%	1787	21.37%		1328	28.8%	2230	25.93%
Negative			271	15.44%	696	8.32%		1156	25.07%	1353	15.73%
Positive		Female	1074	67.55%	2457	65.59%	Female	1563	65.45%	2529	52.1%
Other			395	24.84%	937	25.01%		557	23.32%	1359	28%
Negative			121	7.61%	353	9.4%		268	11.22%	966	19.9%
			POC		White			POC		White	
Positive	Tenured	Male	2567	63.29%	6151	66.31%	Male	3333	50.71%	9790	61.7%
Other			982	24.21%	2242	24.17%		1776	27.02%	4231	26.67%
Negative			507	12.5%	883	9.52%		1464	22.27%	1846	11.63%
Positive		Female	2807	61.27%	1847	63.67%	Female	919	65.69%	3348	63.48%
Other			1208	26.37%	745	25.68%		339	24.23%	1343	25.46%
Negative			566	12.36%	309	10.65%		141	10.08%	583	11.05%

Table 2. Classification Table of UCSC Labels

4.2 Multinomial Logistic Regression

With the three outcomes for label, the regressions have two logit equations, with the negative label used as the reference label for each. More specifically, letting *x* denote the vector of binary encoded covariates along with the two-way, three-way, and four-way interactions of the covariates. Denote the probability of a positive, negative, and other label as *p*, *n*, and *o*, respectively. The two logit equations are,

$$\text{Logit 1: } \log\left(\frac{o}{n}\right) = \alpha_1 + \beta_1' x \quad , \quad \text{Logit 2: } \log\left(\frac{p}{n}\right) = \alpha_2 + \beta_2' x \, .$$

Taken together, the two logit equations imply the probabilities of the label outcomes are,

$$\Pr\text{ (other label)} = \frac{\exp(\alpha_1 + \beta_1' x)}{1 + \exp(\alpha_1 + \beta_1' x) + \exp(\alpha_2 + \beta_2' x)}$$

$$\Pr\text{ (positive label)} = \frac{\exp(\alpha_2 + \beta_2' x)}{1 + \exp(\alpha_1 + \beta_1' x) + \exp(\alpha_2 + \beta_2' x)}$$

$$\Pr\text{ (negative label)} = \frac{1}{1 + \exp(\alpha_1 + \beta_1' x) + \exp(\alpha_2 + \beta_2' x)} \, .$$

Estimates of α_1, β_1, α_2, and β_2 for the UCR data is shown in Table 3. Estimated coefficients for effects not shown (e.g., Female, non-White, non-Tenured, non-STEM, and any interactions that involve these levels) are zero. The yellow highlighted cells identify which p-values are smaller than the traditional threshold of 0.05. It can be seen that overall, and particularly for the second logit equation, all four covariates are involved in the statistically significant p-values.

Estimates of α_1 , β_1 , α_2 , and β_2 for the UCSC data are similarly shown in Table 4. It can be seen that compared to the UCR data analysis, more coefficients are significant across both logit equations and that the standard errors of the estimates are smaller. Both of these observations are the result of the substantially larger sample size for the UCSC dataset.

	Logit 1: Other relative to N			Logit 2: Positive relative to N		
Effect	Estimate	Std. Error	P-Value	Estimate	Std. Error	P-Value
Intercept	0.251	0.23	.26	1.786	0.18	0
Male (M)	0.140	0.27	.61	0.795	0.22	< .001
White (W)	0.443	1.25	.72	3.435	1.02	< .001
Tenured (T)	-0.272	0.27	.31	0.344	0.21	.10
STEM (S)	0.518	0.36	.15	0.658	0.30	.028
M x W	-0.515	1.26	.68	-4.411	1.03	< .001
M x T	0.716	0.34	.036	0.134	0.28	.63
M x S	-0.259	0.41	.53	-1.464	0.34	< .001
W x T	0.358	1.26	.78	-3.685	1.03	< .001
W x S	0.265	1.31	.84	-2.558	1.08	.018
T x S	-0.818	0.42	.050	-1.686	0.34	< .001
W x T x S	-0.266	1.34	.84	3.189	1.11	.004
M x T x S	-0.122	0.49	.80	0.494	0.40	.22
M x W x S	-0.329	1.33	.81	2.926	1.10	.008
M x W x T	-0.472	1.29	.71	4.858	1.06	< .001
M x W x T x S	0.502	1.38	.72	-3.687	1.14	.001

Table 3. Multinomial Logistic Regression Model for UCR Data

	Logit 1: Other relative to N			Logit 2: Positive relative to N		
Effect	Estimate	Std. Error	P-Value	Estimate	Std. Error	P-Value
Intercept	1.183	0.10	0	2.183	.10	0
Male (M)	-0.654	0.13	< .001	-0.854	.12	< .001
White (W)	-0.240	0.11	.034	-0.0495	.10	.63
Tenured (T)	-0.425	0.12	< .001	-0.582	.11	< .001
STEM (S)	-0.452	0.13	< .001	-0.420	.12	< .001
M x W	0.690	0.15	< .001	0.663	.14	< .001
M x T	0.557	0.15	< .001	0.875	.14	< .001
M x S	0.0613	0.15	.69	-0.299	.14	.032
W x T	0.362	0.14	.010	-/236	.13	.068
W x S	-0.150	0.14	.29	-0.751	.13	< .001
T x S	0.571	0.17	< .001	0.693	.15	< .001
W x T x S	-0.0147	0.20	.94	0.438	.18	< .001

M x T x S	-0.648	0.20	.001	-0.773	.18	< .001
M x W x S	0.0608	0.18	.74	0.838	.16	< .001
M x W x T	-0.542	0.19	.003	-0.531	.17	.002
M x W x T x S	0.469	0.24	.053	0.00159	.22	.99

Table 4. Multinomial Logistic Regression Model for UCSC Data

4.3 Clustering Analysis

Because the probabilities for the three labels sum to one, the estimated trinomial distributions were plotted as individual points within a two-dimensional simplex. Figure 1 shows this plot for the UCR data. Each of the 16 estimated trinomial distributions is represented by the combination of one of four symbols used for ethnicity and gender features (circle, triangle, square, cross) and one of four colors used for tenure status and discipline area features (red, green, blue, purple).

A K-means algorithm in combination with silhouette plots (Rousseeuw, 1987) found two clusters in Figure 1, with the seven points in the upper left portion of the plot corresponding to one cluster and the nine points in the lower right portion of the corner corresponding to the second cluster. The first cluster corresponds to distributions that have a relatively high probability of positive comments and a relatively low probability of negative comments, while the second cluster has the converse.

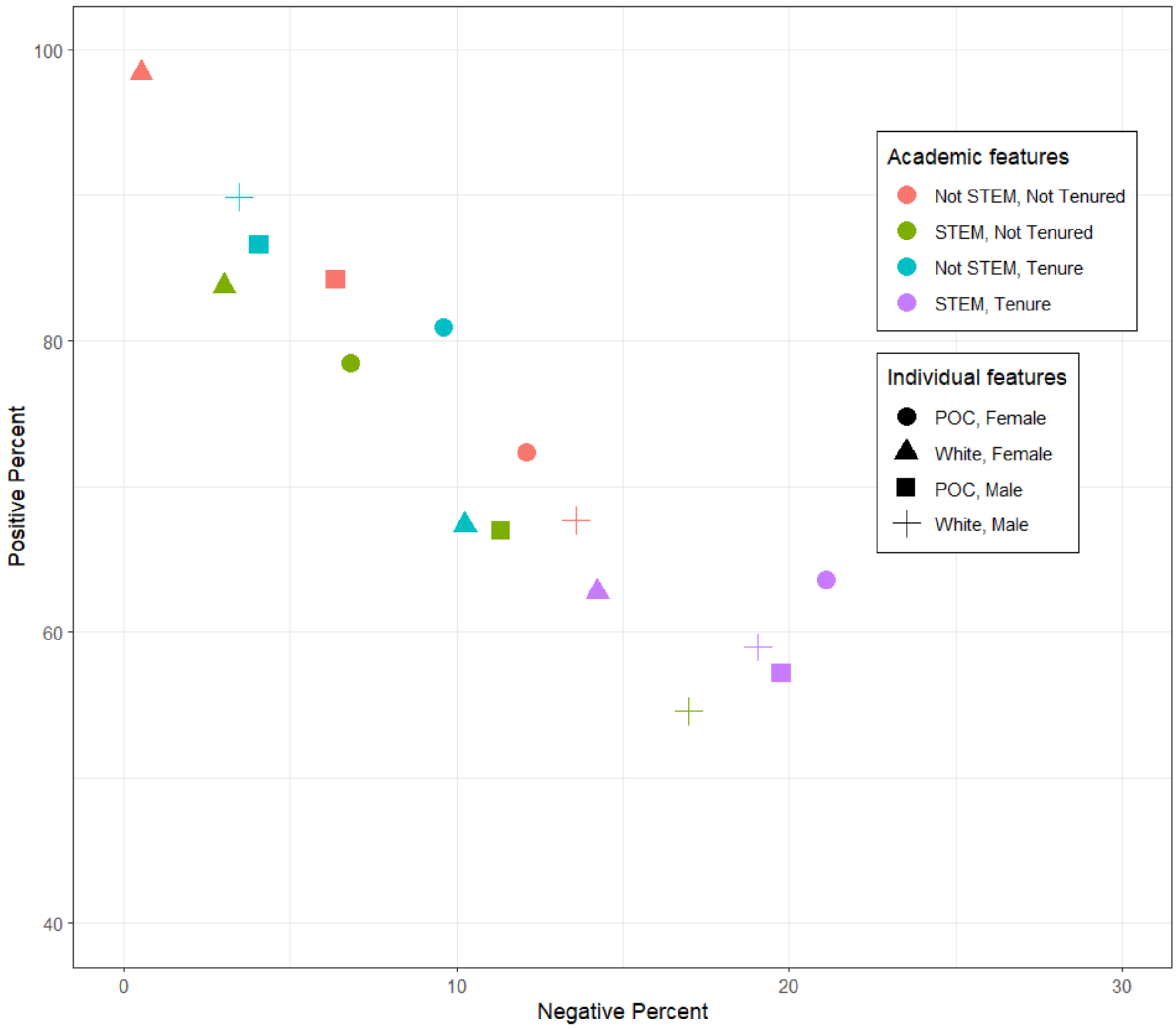

Figure 1. Representation of the Estimated Trinomial Distributions for the UCR Data

Figure 2 is the simplex plot for the UCSC data. Again the K-means algorithm and use of the silhouette plots suggests two clusters of distributions. The upper left portion of the plot, containing eleven distributions, is the first cluster, and the lower right portion of the plot containing five distributions is the second cluster. The yellow shading in Table 2 shows the distributions that fall into the first cluster.

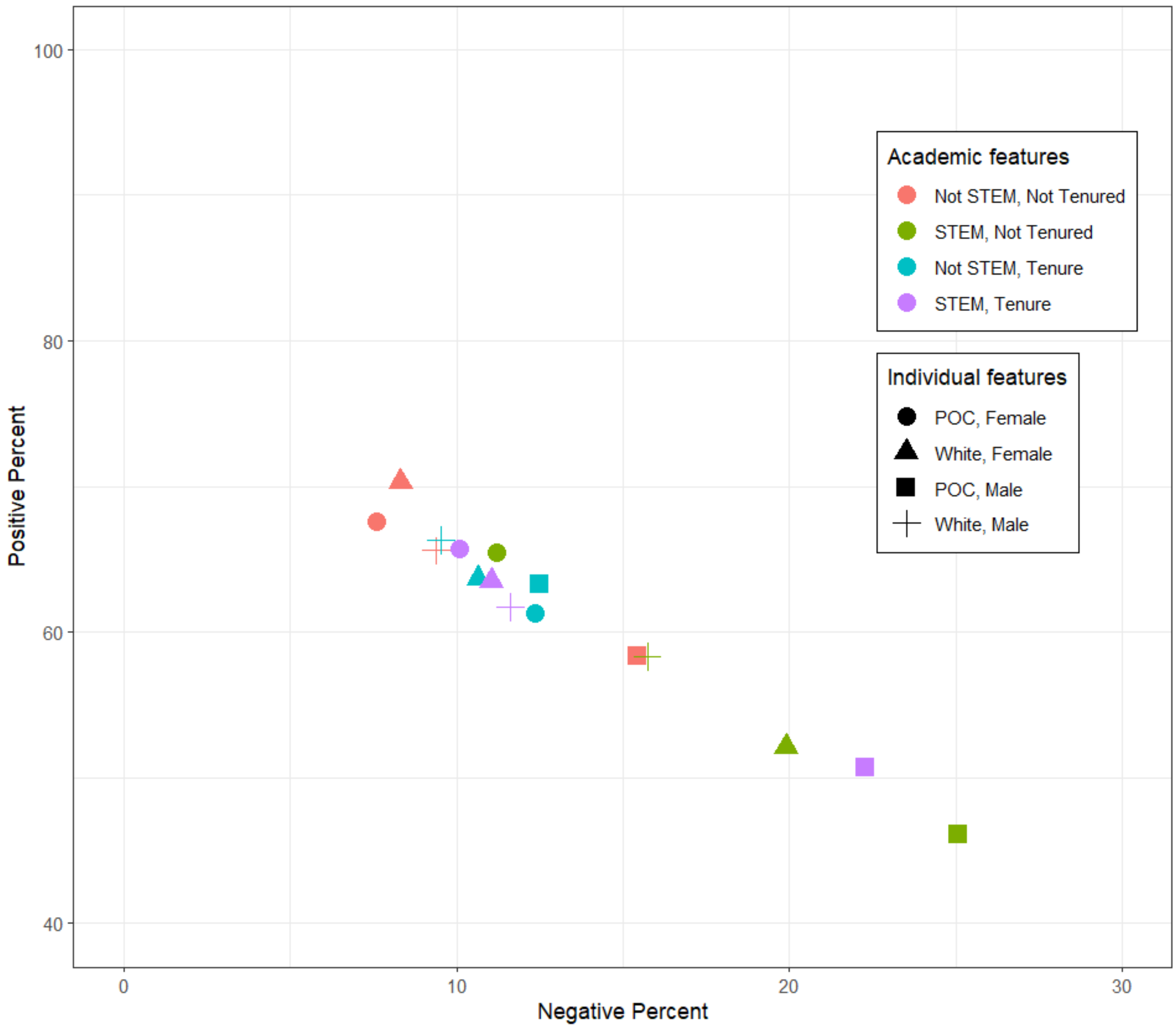

Figure 2. Representation of the Estimated Trinomial Distributions for the UCSC Data

5. Findings

Yellow shading of the trinomial distributions in Tables 1 and 2 correspond to distributions that (from the clustering analysis) have a relatively high probability of a positive comment and a relatively low probability of negative comments. Similarly, trinomial distributions not shaded correspond to distributions that have a relatively low probability of a positive comment and a relatively high probability of negative comments. Any of the trinomial distributions shaded yellow have no practical difference with respect to each other, and likewise, any of the trinomial distributions not shaded have no practical difference with respect to each other. However, any trinomial distributions that differ in how they are shaded have a practical difference. Of primary interest to us was assessing the effects on how comments are labeled by gender and ethnicity within the combinations of tenure status and discipline area. To that end, Table 5 summarizes and contrasts the conclusions for each of those combinations and for each datasets.

Referring to the clusters for UCR data in Figure 1, the upper left cluster that captures distributions with a relatively high probability of positive comments and a relatively low probability of negative comments shows balance in the distributions that represent gender, four female distributions versus three male distributions, and the distributions that represent ethnicity four POC distributions versus three white distributions. Referring to the same cluster for the UCSC data in Figure 2, there is imbalance with respect to the distributions that represent gender, with seven female distributions versus four male distributions, but there is balance with respect to the distributions that represent ethnicity, with five POC distributions versus six white distributions.

Examining the lower right cluster that captures distributions with a relatively low probability of positive comments and a relatively high probability of negative comments, the UCR data in Figure 1 shows balance in the distributions that represent gender, with four female distributions versus five male distributions, and the distributions that represent ethnicity, with four POC distributions versus five white distributions. Referring to the same cluster for the UCSC data in Figure 2, there is imbalance with respect to the distributions that represent gender, with one female distribution versus four male distributions, and balance with respect to the distributions that represent ethnicity, three POC distributions versus two white distributions.

Tenure Status	Course Type	UCR	UCSC
Tenured	STEM	No practically significant ethnicity effect was identified. No practically significant gender effect was identified.	Within female instructors, no practically significant ethnicity effect was identified, but within male instructors, white instructors received a significantly higher proportion of positive comments compared to POC instructors. Within white instructors no practically significant gender effect was identified, but within POC instructors female instructors received a significantly higher proportion of positive comments than male instructors.
Not	STEM	No practically significant ethnicity effect	Within male instructors no practically

Tenured		was identified. Female instructors received a significantly higher proportion of positive comments than male instructors.	significant ethnicity effect was identified, but within female instructors, POC instructors received a higher proportion of positive comments than white instructors. Within white instructors, no practically significant effect was identified, but within POC instructors female instructors received a significantly higher proportion of positive comments than male instructors.
Tenured	Not STEM	Within male instructors there was no practically significant ethnicity effect identified, but within female instructors POC instructors received a significantly higher proportion of positive comments than white instructors. Within POC instructors there was no practically significant gender effect, but within white instructors male instructors received a significantly higher proportion of positive comments than female instructors.	No practically significant ethnicity effect was identified. No practically significant gender effect was identified.
Not Tenured	Not STEM	Within male instructors, POC instructors received a significantly higher proportion of positive comments than white instructors, but within female instructors, white instructors received a significantly higher proportion of positive comments than POC instructors. Within POC instructors, male instructors received a significantly higher proportion of positive comments than female instructors, but within white instructors, female instructors received a significantly higher proportion of positive comments than male instructors.	Within female instructors, no practically significant ethnicity effect was identified, but within male instructors white instructors received a significantly higher proportion of positive comments than POC instructors. Within white instructors there was no practically significant gender effect identified, but within POC instructors female instructors received a significantly higher proportion of positive comments than male instructors.

Table 5. Summary and Comparisons of Findings

6. Summary

While there were combinations of the factors studied (tenure status, discipline area, gender, and ethnicity) that advantaged instructors with respect to having a higher probability of receiving positive comments, there was no evidence that advantages skewed in a consistent way. Consider the gender comparison for the eight combinations of tenure status, discipline area, and ethnicity. In the UCR data, three of these comparisons point to female instructors having an advantage, two of the combinations point to male instructors having an advantage, and there was no practical difference identified in the other three combinations. The primary finding appears to be a disadvantage for tenured instructors of STEM classes, which are all in the lower cluster, regardless of individual demographic features. In the UCSC data, three of the combinations pointed to female instructors having an advantage, with no practical difference in the other five combinations. Non-tenured instructors of STEM classes appear to have a disadvantage, primarily appearing in the lower cluster.

Next, consider the ethnicity comparison for the eight combinations of tenure status, discipline area, and gender. For the UCR data, two of the combinations pointed to POC instructors having an advantage, one combination pointed to white instructors having an advantage, and there was no practical difference in the other five combinations. In the UCSC data, one combination pointed to POC instructors having the advantage, two combinations pointed to white instructors having an advantage, and there was no practical difference in the other five combinations.

Finally, the two campuses differed in important ways. First, by comparing Figure 1 to Figure 2, it can be seen that overall there was a higher proportion of positive comments at UCR. Second, by examining Table 5 the conclusions where significant practical effects were found differ substantially between the two campuses. For example, there was no gender or ethnicity effect at UCR for tenured instructors teaching STEM classes, whereas there was at UCSC. On the other hand, there was no gender or ethnicity effect at UCSC for tenured instructors teaching non-STEM classes, whereas there was at UCR.

At UCR, the lack of consistent bias by gender or ethnicity supports the perspective that bias in student narrative evaluations of teaching does not consistently disadvantage instructors of a particular gender or ethnicity. To the extent that bias exists, it seems more likely to derive from the context of the instruction, which includes the subject being taught, the teaching environment, the approach taken by the instructor, and the attitudes of the student. In contrast, at UCSC, there was a higher proportion of negative narrative comments for instructors who were male and POC. There is a need for additional study to understand why male POC instructors tend to have worse narrative comments. One possibility is that negative comments are correlated with stronger accents, consistent with the findings of Subtirelu (2015) and Wang and Gonzalez (2020). Within STEM fields, UCSC POC faculty are predominantly Asian. An unrelated survey of instructors of STEM classes found that Asian male faculty are more likely to be immigrants and more likely to self-report a noticeable accent than other groups, including Asian female STEM faculty.

It is striking that on neither campus was there evidence of lower evaluations for female instructors, which is not consistent with the larger literature on course evaluations that usually find bias against female instructors. It is not clear if there is something different about narrative comments compared to the multiple-choice or ranked questions typically studied in the course evaluation literature, and this is a topic that deserves further investigation. Overall, it appears that the narrative comments show less bias than traditional multiple-choice questions, and thus may be more appropriate for use in evaluation of instructors.

Acknowledgements

This work was supported by an Advancing Faculty Diversity Grant from the University of California Office of the President.

References

Adams, Sophie, Bekker, Sheree, Fan, Yanan, Gordon, Tess, Shepherd, Laura J., Slavich, Eve, and David Waters. 2022. "Gender Bias in Student Evaluations of Teaching: Punish[ing] Those Who Fail To Do Their Gender Right," *Higher Education*, 83:787-807.

Agresti, A. 2007. *An Introduction to Categorical Data Analysis*, 2nd ed., New York: John Wiley & Sons.

Basow, Susan A. and Montgomery, Suzanne. 2005. "Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation," *Journal of Personnel Evaluation in Education*, 18: 91-106.

Boring, Anne, Ottoboni, Kellie, and Stark, Philip B. (2016). "Student evaluations of teaching (mostly) do not measure teaching effectiveness," *ScienceOpen Research*, DOI: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1.

Carrell, Scott E. and West, James E. (2010). "Does professor quality matter? Evidence from random assignment of students to professors," *Journal of Political Economy*, 118: 409-432.

Chávez, Kerry and Mitchell, Kristina M. W. 2020. "Exploring bias in student evaluations: Gender, race, and ethnicity," *PS: Political Science & Politics*, 53(2): 270-274.

Ceci, Stephen, J., Shulamit, Kahn, and Williams, Wendy M. 2023. "Exploring Gender Bias in Six Key Domains of Academic Science: An Adversarial Collaboration," *Psychological Science in the Public Interest*, 24(1): 15-73.

El-Alayli, Amani, Hansen-Brown, Ashley A., and Ceynar, Michelle. (2018). "Dancing backwards in high heels: Female professors experience more work demands and special favor requests, particularly from academically entitled students," *Sex Roles*, 79(3-4), 136-150.

Fan, Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., and Johnston, E. L. (2019). "Gender and cultural bias in student evaluations: Why representation matters," *PLoS One*, 14(2), e0209749.

Gelber, Katharine, Brennan, Katie, Duriesmith, David and Fenton, Ellyse. 2022. "Gendered mundanities: Gender bias in student evaluations of teaching in political science," *Australian Journal of Political Science*, 57(2): 199-220.

Lindahl, Mary W. and Unger, Michael L. 2010. "Cruelty in Student Teaching Evaluations," *College Teaching*, 58 (3): 71–76.

Littleford, Linh N., Ong, Katherine S., Tseng, Andy, Milliken, Jennifer C., and Humy, Sonya L. 2010. "Perceptions of European American and African American Instructors Teaching Race-Focused Courses," *Journal of Diversity in Higher Education*, 3 (4):230–44

MacNell, Lillian, Driscoll, Adam, and Hunt, Andrea N. 2015. “What’s in a name: Exposing gender bias in student ratings of teaching,” *Innovative Higher Education*, 40: 291-303.

Mengel, Friederike, Sauermann, Jan, and Zölitz, Ulf. 2018. “Gender bias in teaching evaluations,” *Journal of the European Economic Association*, 17(2): 535–566.

Mitchell, Kristina and Martin, Jonathan. 2018 “Gender bias in Student Evaluations,” *PS: Political Science & Politics*, 51(3): 648-652.

Reid, Landon D. (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), 137-152.

Rousseeuw, P. J. 1987 “Silhouettes: a Graphical Aid to the Interpretation and Validation of cluster Analysis,” *Computational and Applied Mathematics*, 20: 53-65.

Schmidt, Benjamin. 2015. “Gender bias exists in professor evaluations,” *New York Times*, December 16, 2015.

Signorell, A. DescTools: Tools for Descriptive Statistics, R package Version 0.99.52 (2023), https://cran.r-project.org/web/packages/DescTools/index.html.

Stoesz, Brenda M., De Jaeger, Amy E., Quesnel, Matthew, Bhojwani, Dimple, and Los, Ryan. (2022). “Bias in Student Ratings of Instruction: A Systematic Review of Research from 2012 to 2021,” *Canadian Journal of Educational Administration and Policy*, 201: 39-62.

Subtirelu, Nicholas C. (2015), “She does have an accent but...: race and language ideology in students’ evaluations of mathematics instructors on RateMyProfessors.com,” *Language in Society*, Vol. 44 (1): 35-62.

Uttl, Bob, White, Carmela A., and Gonzalez, Daniela W. (2017). “Meta-analysis of faculty’s teaching effectiveness: Student evaluation of teaching ratings and student learning are not related,” *Studies in Educational Evaluation*, 54: 22-42.

Wagner, Natascha, Rieger, Matthias, and Voorvelt, Katherine. 2016. “Gender, ethnicity and teaching evaluations: Evidence from mixed teaching teams,” *Economics of Education Review*, 54: 79–94.

Wallace, Sherri L., Lewis, Angela K., and Allen, Marcus D. 2019. “The State of the Literature on Student Evaluations of Teaching and an Exploratory Analysis of Written Comments: Who Benefits Most?” *College Teaching*, 67 (1): 1–14.

Wang, Lei, and Gonzalez, Jorge A. (2020). “Racial/ethnic and national origin bias in SET,” *International Journal of Organizational Analysis, 28*(4), 843–855.