Gender Biases in TA Evaluation

Radhika Bansal

University of California Riverside

Interim Draft

This Version, March 23, 2026

Abstract

This study examines whether gender bias influences how students evaluate their teaching assistants (TAs). Using administrative, individual-level data on freshman students and TAs at a large public university in California from Fall 2016 to Summer 2025, we test whether students assign different evaluations based on gender match. Leveraging within-course variation in TA assignment, we isolate the causal effect of student–TA gender congruence while holding course content and structure constant.

We find evidence of a statistically significant same-gender effect in teaching evaluations. Female students rate female TAs approximately 0.17 standard deviations higher relative to male students rating male TAs. The results are robust when focusing on the fall quarter of freshman year, excluding COVID-affected terms, and restricting the sample to courses with a single TA. These findings suggest that gender congruence meaningfully shapes student evaluations and highlight the role of in-group bias in academic assessment contexts.

1 Introduction

A growing body of literature documents systematic gender bias in student evaluations of teaching (SETs), particularly against female instructors and teaching assistants (TAs). Students often penalize female TAs when they do not conform to traditional feminine norms. Using data from a French university, Boring (2017) shows that male students exhibit a bias in favor of male instructors. Moreover, the dimensions along which teaching is evaluated differ by instructor gender and align with prevailing stereotypes: male instructors are perceived by both male and female students as more knowledgeable and as possessing stronger classroom leadership skills.

Consistent with these findings, Mengel et al. (2019) demonstrates that women receive systematically lower teaching evaluations than their male counterparts. This bias is primarily driven by male students, is more pronounced in mathematically intensive courses, and is especially strong for junior female instructors. Together, these results suggest that both the level and structure of teaching evaluations are shaped by gendered perceptions of competence and authority.

Experimental evidence further reinforces the presence of such biases. In a study by Khazan et al. (2019), students were randomly assigned to a putative male or female TA, although all tasks were performed by the same female TA. Despite the absence of direct interaction in the online environment, evaluations differed significantly: the TA presented as female received substantially more negative evaluations than when presented as male. Interestingly, the largest discrepancies emerged in evaluations of the putative female TA, with particularly low ratings often given by female students. However, not all experimental evidence points in the same direction. Andersson et al. (2023) conducts a double-blind experiment in which male or female names were randomly assigned to online instructors. Students evaluated mentors on helpfulness, knowledge, and responsiveness, and the results show no

evidence of bias against female instructors across these dimensions. This contrast highlights that the extent of gender bias may depend on context, evaluation criteria, and the nature of student–instructor interaction. A similar experiment by (**?**) where the same instructor in an online class each operated under two different gender identities. Students rated the male identity significantly higher than the female identity, regardless of the instructor's actual gender.

Recent work also emphasizes the role of sorting and equilibrium effects in shaping observed biases. Ayllón et al. (2025) argues that gender bias in teaching evaluations is mediated by students' ability to sort into institutions, majors, courses, and instructors based on their preferences. While female students exhibit significantly less bias against female faculty than male students, most variation in gender bias remains idiosyncratic at the individual level. Notably, female students tend to sort toward female instructors both across and within fields, suggesting that preferences over instructor gender can influence academic trajectories.

Beyond evaluations, a related literature examines how instructor characteristics shape students' academic choices. Prior research shows that students tend to form stronger academic relationships with instructors or TAs who share similar backgrounds, such as ethnicity. Additionally, exposure to role models plays an important role in shaping educational pathways: for instance, students are more likely to choose STEM majors when advised by faculty affiliated with STEM fields, and female students are more likely to pursue STEM when they have female advisors with STEM backgrounds Canaan and Mouganie (2023).

This study contributes to the literature by examining whether gender bias shapes both student evaluations and instructor-related outcomes using data from an entire U.S. university, rather than focusing on specific departments, thereby capturing substantial institutional diversity in terms of gender and field of study. Specifically, it analyzes whether female students evaluate female TAs or instructors more favorably than male instructors in otherwise identical settings. Moreover, we try to see if the own-group bias is actually relative bias or

statistical discrimination. Additionally, we extend the analysis to assess whether observed own-gender bias reflects a broader form of in-group bias **?**. In particular, we examine whether TAs and students from similar identity groups share characteristics beyond gender—many of which may be unobserved—and whether students from different identity groups exhibit correlated preferences or biases over these characteristics.

2 Background and Data

2.1 Institutional Setup

We draw on administrative data from a large public university in California, a large public university. UCR operates on a quarter system consisting of ten weeks of instruction followed by a final examination week. The academic year begins in the fall and continues through the winter and spring quarters, with an additional summer term composed of shorter instructional sessions.

Course enrollment follows a decentralized process in which students register themselves into lectures and, where applicable, associated discussion sections. Enrollment typically occurs prior to the start of the quarter and continues through the end of Week 2, subject to seat availability. At the time of registration, students can most certainly observe the identity of the primary instructor for the course but may not be informed about the teaching assistants (TAs) assigned to discussion sections. As course instructors are determined way in advance but TA assignments might happen till very last before the quarter begins **confirm**. Not all courses include a discussion component.

For courses that do include discussions, these sections meet once per week for approximately 50 minutes and have a class strength limit of 30 students. Instructors in these discussions can be graduate students.**CONFIRM** The structure and content of discussion sections are determined by the primary instructor and can vary across courses and instruc-

tors. Common uses include reviewing lecture material, working through additional practice problems, and discussing homework assignments. Importantly, participation in discussion sections does not typically carry additional or separate course credit, making them a complementary instructional component rather than a graded requirement.

Student evaluations of teaching are administered systematically across terms but follow slightly different schedules depending on the quarter. During the fall, winter, and spring quarters, students complete evaluations of both their teaching assistants (TAs) and primary instructors during weeks 9 and 10 of instruction, just prior to final examinations. In the summer term, evaluations are administered over a shorter window: the iEval system is open during the final four weekdays (Monday through Thursday) of each session, and submissions close before the start of final exams. Importantly, to preserve the integrity of grading, both TAs and instructors are granted access to evaluation results only after final grades have been submitted.

3 A Simple Model of Student Evaluations

Consider a setting where student i evaluates instructor (or TA) j. The observed evaluation score is given by:

$$Y_{ij} = Q_j - C_{ij} + B_{ij} + \varepsilon_{ij}, \tag{1}$$

where Q_j denotes the true teaching quality of instructor j, C_{ij} captures mismatch or evaluation costs between student i and instructor j, B_{ij} represents bias or preference components, and ε_{ij} is an idiosyncratic error term.

Let F_i and F_j be indicator variables equal to one if the student and instructor are female, respectively. We model the bias component as:

$$B_{ij} = \alpha_1 F_i + \alpha_2 F_j + \alpha_3 (F_i \times F_j), \tag{2}$$

where α_1 captures baseline differences in evaluation behavior across student gender, α_2 captures the average effect of instructor gender, and α_3 captures own-gender bias.

We allow the mismatch term to depend on unobserved characteristics:

$$C_{ij} = \gamma \cdot \text{Mismatch}_{ij}, \tag{3}$$

where Mismatch_{ij} may reflect differences in communication styles, expectations, or other unobserved traits that can be correlated with gender.

Substituting, the evaluation equation becomes:

$$Y_{ij} = \alpha_1 F_i + \alpha_2 F_j + \alpha_3 (F_i \times F_j) + Q_j - C_{ij} + \varepsilon_{ij}. \tag{4}$$

After controlling for instructor quality and other observables, the empirical specification can be written as:

$$Y_{ij} = \beta_1 F_i + \beta_2 F_j + \beta_3 (F_i \times F_j) + u_{ij}, \tag{5}$$

where u_{ij} captures unobserved determinants of evaluations.

The coefficient β_3 is typically interpreted as evidence of own-gender bias. However, it may also capture match-specific effects arising from unobserved characteristics that are correlated with gender. As a result, β_3 does not, on its own, uniquely identify discriminatory preferences.

3.1 Sample Construction

The dataset contains rich, de-identified student-level information. This includes demographic characteristics such as gender, race, and ethnicity, as well as pre-college academic preparation measured by high school GPA. We also observe detailed academic records, including course enrollments by term, term GPA, and declared major. These student records are linked to

de-identified teaching assistants, allowing us to study interactions between students and TAs within specific courses and terms.

In addition to administrative records, the data include quantitative evaluation scores submitted by students anonymously for each course in each term. The students answer different questionnaires for TAs and the primary instructors. **ADD THE REFERENCE TO QUESTIONNAIRE**

These evaluations provide standardized measures of student perceptions of both TAs and instructors, enabling us to analyze how instructional characteristics and student–TA interactions relate to teaching effectiveness and student outcomes over time.

We restrict our sample to first-time freshman students using data spanning Fall 2016 through Summer 2025, which represents the earliest period for which consistent data are available. We end the sample in Summer 2025 because the university transitioned to a new evaluation system beginning in Fall 2025, making responses not directly comparable across periods.

Focusing on first-time freshmen helps mitigate concerns related to endogenous course selection, as these students typically face more structured and constrained enrollment choices in their initial terms compared to more advanced students.

As an additional robustness check, we further restrict the sample to students in the fall quarter of their freshman year, when course choice is most limited, to ensure that our results are not driven by residual selection into courses or sections.

In addition, we limit the sample to students who self-identify as either male or female, as our analysis focuses on gender-based interactions and outcomes.

3.2 Evaluations

Students complete separate evaluation forms for teaching assistants (TAs) and primary instructors anonymously. In our analysis, we focus exclusively on the TA evaluation forms.

These forms consist primarily of a set of quantitative questions, along with an open-ended qualitative component that allows students to provide written feedback. While the exact set of questions varies slightly between discussion and lab sections—with some items specific to each format—there is a common subset of questions shared across both. To ensure consistency, we restrict our analysis to the 10 overlapping questions.

The quantitative section includes multiple items designed to capture different dimensions of TA effectiveness in discussion sections. These questions ask students to evaluate aspects such as the TA's clarity and organization, subject knowledge, ability to explain material and connect it to course content, encouragement of student participation, and the usefulness of feedback provided on assignments. Additional questions assess interpersonal and classroom-related attributes, including approachability, accessibility outside of class, respect for diverse perspectives, and the ability to foster an inclusive and supportive learning environment. Students also provide an overall assessment of the TA's teaching effectiveness. All responses are recorded on a Likert scale ranging from 1 to 7, with higher values indicating more favorable evaluations. This standardized scale allows for consistent comparison of TA performance across courses, sections, and terms.

We construct a composite student evaluation index using principal component analysis on ten standardized evaluation items. The first principal component explains 69.2% of the total variance and loads positively and similarly on all items, indicating a single dominant latent factor.[1]

We group individual evaluation items into broader conceptual categories based on the underlying dimension of teaching they are intended to capture. Using the content of each question, we construct three composite indices, each defined as the row-wise mean of standardized responses to a set of related items.

The first index, Response to Diversity and Classroom Climate, captures the extent to

[1]Eigenvalue = 6.92 : PC1 explains as much variance as 7 original variables. No item dominates the index No item behaves oddly

which the TA fosters an inclusive and respectful learning environment. It combines measures of how appropriately the TA handles sensitive situations or inappropriate student behavior, as well as whether they encourage and respect diverse perspectives in the classroom.

The second index, Instructor Approachability, reflects the accessibility and interpersonal engagement of the TA. This measure aggregates responses on whether the TA encourages student participation, is available and helpful outside of class, maintains an approachable demeanor, and motivates students to perform to their potential.

The third index, Teaching Effectiveness, summarizes core instructional quality. It includes evaluations of the TA's preparedness and organization, clarity and appropriateness of language, overall effectiveness as an instructor, and ability to create and sustain a supportive learning environment. By constructing these indices as averages of standardized responses, we reduce measurement noise from individual survey items and obtain more stable, interpretable measures of distinct dimensions of TA performance.

All quantitative responses are recorded on a standardized rating scale, which enables consistent comparison of TA performance across courses, terms, and student groups.

3.3 Dealing with Selection and Missingness Issues

Given our setting, several sources of selection and sorting may arise. First, while we observe all students enrolled in a given quarter regardless of whether they submit evaluations, key demographic information is sometimes missing. In particular, approximately 14% of students in the raw data lack reported gender. This issue is more pronounced for teaching assistants (TAs), who are themselves students, with about 58% missing gender information.

To address this missingness, we supplement our data by scrapping instructor and TA names for all courses offered at UCR during our sample period from the university's course listings website (classes.ucr.edu). We then merge these records with our discussion-course data at the term level. Using the resulting name information, we infer gender through the

web-based API service Genderize.io, which predicts the likely gender associated with a given first name. This service draws on a large database constructed from over 900 million names across 249 countries, allowing us to systematically impute missing gender information for both students and TAs. [2]. To ensure reliability, we restrict the imputation to cases where the predicted probability of the assigned gender exceeds 80%. We present results using both the original data and the dataset with imputed gender to assess robustness. Following this procedure, the share of missing TA gender declines substantially (to approximately 12%, subject to final verification).

In addition, we construct a third measure of TA gender in which we use the original gender information of TAs but replace the remaining missing data with the imputed gender data, allowing us to further test the sensitivity of our results to assumptions about gender classification.

We formally test if the missingness in TA gender is jointly related with student demographics for all three measures and find no joint significance.[REWORD]

Second, because submitting evaluations is voluntary, there is a potential concern about selection into response. In particular, students who choose to complete evaluations may differ systematically from those who do not, which could bias our estimates.

To assess this concern, we examine whether observable student characteristics predict the likelihood of submitting an evaluation. Specifically, we estimate regressions of an indicator for evaluation submission on student-level observables. This approach allows us to test whether response behavior is systematically related to observable characteristics and, consequently, to gauge the extent of potential selection bias in our sample.

The results are reported in Table ??. While some observable characteristics predict whether a student submits an evaluation, the absence of a significant interaction effect implies that differential response behavior is not systematically driven by the gender match

[2]https://genderize.io/

between students and TAs and TA gender is not significantly correlated with the response behavior of male students.

This is important because the key parameter of interest relies on variation in student–TA gender interactions. If students were more (or less) likely to submit evaluations depending on whether they were matched with a TA of the same or different gender, the observed sample of evaluations could be selectively composed along precisely this dimension, biasing our estimates.

The lack of a significant interaction suggests that, although selection into response exists, it does not operate in a way that is correlated with the main interaction of interest. As a result, concerns that our findings are driven by differential survey participation across student–TA gender pairings are substantially mitigated. **REWORD**

Finally, because students self-select into courses, there is a potential concern about sorting across sections. For instance, female students may disproportionately enroll in courses taught by female TAs, or students of certain backgrounds may sort based on TA gender. Similarly, more motivated students or those expecting higher grades might select into courses associated with a particular TA gender.

We address this concern in two ways. First, we restrict the sample to first-time freshman students, who have limited information at the time of enrollment. This substantially reduces the scope for sorting based on TA characteristics.

Second, we formally test for systematic sorting by examining whether observable student characteristics predict TA gender assignment. Specifically, we conduct a joint test of student characteristics in a regression of an indicator for TA gender. The results show no evidence of systematic sorting: the joint F-statistic is 1.113, indicating that student characteristics are not jointly significant predictors of TA gender.

4 Econometric Strategy

We estimate the following regression equation:

$$Eval_{stcqy} = \beta_0 + \beta_1 g_s + \beta_2 g_t + \beta_3 (g_s \times g_t) + X_s'\gamma + Z_{sqy}'\theta + \phi_c + \delta_{qy} + \varepsilon_{stcqy} \tag{6}$$

where $Eval_{stcqy}$ denotes the standardized principal component of the evaluation provided by student s to teaching assistant t in course c during quarter q of academic year y. The variable g_s is an indicator equal to one if the student is female, and g_t is an indicator equal to one if the TA is female. The vector X_s includes student-specific, time-invariant characteristics such as ethnicity and high school GPA. In addition, we control for the student's academic performance in the given term, measured by term GPA, in vector Z_{sqy}' to account for differences in ability or effort that may influence evaluations. We include course fixed effects, ϕ_c, to absorb time-invariant differences across courses, and quarter-by-year (term) fixed effects, δ_{qy}, to capture common shocks or trends across academic terms. Standard errors are clustered at the TA level to account for correlation in evaluations across students taught by the same TA.

5 Results

The dependent variable is standardized principal component of evaluations, so all coefficients can be interpreted in standard deviation units. Across specifications, female students assign higher ratings than male students, with estimates ranging from 0.06 to 0.08 standard deviations, suggesting that female students are, on average, more generous evaluators. In contrast, the coefficient on instructor gender is small and statistically insignificant, indicating that male students do not systematically rate female instructors differently from male instructors.

The interaction between female student and female instructor is positive, large, and pre-

cisely estimated across all specifications, with magnitudes between 0.12 and 0.18 standard deviations. This provides strong evidence of own-gender bias: female students evaluate female instructors significantly more favorably relative to the baseline of male students evaluating male instructors.

Consistent with this, the combined effect of being a female student evaluating a female instructor over a male instructor ranges from 0.11 to 0.16 standard deviations and is statistically significant in most specifications. These results are stable across the inclusion of covariates and sample restrictions, underscoring the robustness of the findings. Overall, the evidence points to a meaningful and systematic own-gender bias in student evaluations, driven primarily by more favorable ratings from female students toward female instructors.

Relating gender bias to gender discrimination. We follow the definitions of monolithic and non-monolithic bias from**?**. In our empirical framework, the estimates reveal evidence of non-monolithic bias in evaluations. Specifically, female students evaluate female TAs significantly more favorably than male TAs ($\beta_2 + \beta_3 > 0$ and statistically significant), whereas male students rate male and female TAs similarly ($\beta_2 \approx 0$ and statistically insignificant). This pattern indicates that differences in evaluations are driven by heterogeneity across student groups rather than a uniform bias.

We also find that female students provide higher ratings overall compared to male students ($\beta_1 > 0$ and statistically significant), including for male TAs which provides strong evidence that the non-monolithic behaviour might not mean that there is prejudice.. This suggests that part of the observed heterogeneity reflects differences in evaluation standards or grading behavior across student groups, rather than purely identity-based favoritism.

Male students appear to act in a more uniform, or monolithic, manner in their evaluations, but this does not necessarily imply the absence of bias. For example, both male and female students could rate male TAs lower, which would produce a monolithic pattern while still

reflecting gender-based preferences.

Importantly, neither β_1 nor β_3 alone is sufficient to establish discriminatory preferences. While β_3 captures differential responses across student groups, β_1 reflects the baseline group's evaluation tendencies, which may themselves be shaped by factors unrelated to gender-based preferences. Differential responses may also arise from variation in expectations, classroom experiences, or other unobserved characteristics correlated with student and instructor gender.

Taken together, these results are consistent with own-gender bias, but should be interpreted as reduced-form evidence of heterogeneous evaluation patterns rather than definitive proof of identity-based favoritism.

However, given that the outcome test is not possible in our setting we move to (Ayllón et al., 2025) and use their setup to identify the gender homophily and relate it with bias. Our setup also satisfies the parallel disparities assumption, which implies that, in the absence of bias, the differences in evaluations between male and female students would be constant across TA gender. Under this assumption, the homophily estimand described by Argyle et al. (2025) identifies the average difference in total gender bias between male and female students. In our regression specification, this estimand corresponds to the interaction term β_3, which captures how female students' evaluations of female TAs differ relative to male students. Thus, β_3 provides a reduced-form measure of student-level homophily or own-gender bias in TA evaluations. In our context, total gender bias in student evaluations includes not only taste-based bias and inaccurate statistical bias (analogous to β_μ-type bias), but also potentially accurate statistical discrimination. Under the Parallel Accurate Statistical Discrimination (PASD) assumption — which posits that any accurate statistical discrimination is similar across male and female students on average — the effect of accurate statistical discrimination does not confound comparisons between student groups. If both the parallel disparities assumption and PASD hold, our homophily estimates (captured by the interaction

term β_3) specifically reflect the contribution of taste-based and inaccurate statistical bias to the observed differences in evaluations across male and female students. However, we cannot calculate bounds or partially identify the $\beta\mu$ gender bias in "bad" evaluations in the same way as Argyle et al. (2025), because the analogous assumption — that, on average, male students are weakly biased against male TAs or in favor of female TAs — does not hold in our setting.

6 Robustness

We conduct several robustness checks to assess the validity of our findings. First, we restrict the sample to students in the fall quarter of their freshman year. These students are least likely to have prior information about TAs, thereby strengthening the assumption of quasi-random assignment between students and TAs.

Second, we re-estimate our main specification using only courses that have more than one TA. This allows us to compare outcomes within courses where multiple TAs are present, further mitigating concerns that our results are driven by course-level differences.

Finally, our sample period overlaps with the COVID-19 pandemic, during which instructional modalities changed substantially. UCR transitioned to fully remote instruction beginning in the final weeks of Winter 2020, with online instruction continuing through Summer 2021. To ensure that our results are not driven by this shift, we exclude quarters from Spring 2020 through Summer 2021 and re-estimate our models. The results remain qualitatively similar, suggesting that our findings are robust to the exclusion of the pandemic period.

7 Heterogeneity

References

Andersson, O., Backman, M., Bengtsson, N., and Engström, P. (2023). Are Economics Students Biased against Female Teachers? Evidence from a Randomized, Double-Blind Natural Field Experiment. Technical report.

Argyle, B., Indarte, S., Iverson, B., Palmer, C., Cahoon, L., Earnest, S., Gillespie, N., Green, B., Jindal, T., Kahley, D., Myzelev, E., Nguyen, S., Petersen, D., Raviola, S., Sivers, E., Sivers, L., Sivers, S., Stevenson, P., Goldsmith-Pinkham, A., Gupta, B., Lewis, F., Liu, E., Mayer, M., Pagel, L., and Wang, E. Y. (2025). NBER WORKING PAPER SERIES RACIAL DISPARITIES AND BIAS IN CONSUMER BANKRUPTCY "Explaining Racial Disparities in Personal Bankruptcy Outcomes." We thank. Technical report.

Ayllón, S., Lefgren, L. J., Patterson, R. W., Nicolás, O. B. S., Andrade, U., Stoddard, O. B., and Andrade, N. U. (2025). NBER WORKING PAPER SERIES 'SORTING' OUT GENDER DISCRIMINATION AND DISADVANTAGE: EVIDENCE FROM STUDENT EVALUATIONS OF TEACHING given to the source. 'Sorting' Out Gender Discrimination and Disadvantage: Evidence from Student Evaluations of Teaching. Technical report.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, 145:27–41.

Canaan, S. and Mouganie, P. (2023). The Impact of Advisor Gender on Female Students' STEM Enrollment and Persistence. *Journal of Human Resources*, 58(2):594–632.

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

Mengel, F., Sauermann, J., and Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association*, 17(2):535–566.

References

Andersson, O., Backman, M., Bengtsson, N., and Engström, P. (2023). Are Economics Students Biased against Female Teachers? Evidence from a Randomized, Double-Blind Natural Field Experiment. Technical report.

Argyle, B., Indarte, S., Iverson, B., Palmer, C., Cahoon, L., Earnest, S., Gillespie, N., Green, B., Jindal, T., Kahley, D., Myzelev, E., Nguyen, S., Petersen, D., Raviola, S., Sivers, E., Sivers, L., Sivers, S., Stevenson, P., Goldsmith-Pinkham, A., Gupta, B., Lewis, F., Liu, E., Mayer, M., Pagel, L., and Wang, E. Y. (2025). NBER WORKING PAPER SERIES RACIAL DISPARITIES AND BIAS IN CONSUMER BANKRUPTCY "Explaining Racial Disparities in Personal Bankruptcy Outcomes." We thank. Technical report.

Ayllón, S., Lefgren, L. J., Patterson, R. W., Nicolás, O. B. S., Andrade, U., Stoddard, O. B., and Andrade, N. U. (2025). NBER WORKING PAPER SERIES 'SORTING' OUT GENDER DISCRIMINATION AND DISADVANTAGE: EVIDENCE FROM STUDENT EVALUATIONS OF TEACHING given to the source. 'Sorting' Out Gender Discrimination and Disadvantage: Evidence from Student Evaluations of Teaching. Technical report.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, 145:27–41.

Canaan, S. and Mouganie, P. (2023). The Impact of Advisor Gender on Female Students' STEM Enrollment and Persistence. *Journal of Human Resources*, 58(2):594–632.

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

Mengel, F., Sauermann, J., and Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association*, 17(2):535–566.

8 Appendix

8.1 Tables

Table 1: Missingness in Response

	response	response	response	response	response	response
Female Student	0.025***	0.025***	0.024***	0.039***	0.028***	0.029***
	(0.002)	(0.002)	(0.002)	(0.005)	(0.004)	(0.004)
Female	-0.006			-0.009		
	(0.005)			(0.011)		
Female		0.003			0.008	
		(0.006)			(0.010)	
Female			-0.001			0.003
			(0.005)			(0.009)
Female Student × Female	0.002			0.003		
	(0.004)			(0.007)		
Female Student × Female		0.000			0.007	
		(0.003)			(0.007)	
Female Student × Female			0.001			0.006
			(0.003)			(0.006)
Student is Asian	-0.005*	-0.003**	-0.003**	-0.000	0.003	0.003
	(0.002)	(0.002)	(0.002)	(0.005)	(0.004)	(0.003)
Student is White	0.000	-0.001	-0.000	0.003	0.001	0.001
	(0.003)	(0.002)	(0.002)	(0.007)	(0.005)	(0.005)
Student is Black	0.001	-0.005	-0.005*	0.015*	0.007	0.005
	(0.004)	(0.003)	(0.003)	(0.009)	(0.006)	(0.006)
Student is Hispanic	0.010***	0.010***	0.010***	0.018***	0.016***	0.017***
	(0.003)	(0.002)	(0.002)	(0.005)	(0.004)	(0.004)
Student_course_Term_GPA	0.032***	0.034***	0.034***	0.042***	0.040***	0.040***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)
High_School_GPA	0.032***	0.031***	0.030***	0.033***	0.023***	0.023***
	(0.003)	(0.002)	(0.002)	(0.005)	(0.004)	(0.004)
Student_current_age	0.004***	0.002	0.002*	0.005	0.003	0.003
	(0.002)	(0.001)	(0.001)	(0.003)	(0.002)	(0.002)
Covariates						
Sample	Freshman	Freshman	Freshman	Freshman, Fall	Freshman, Fall	Freshman, Fall
gender	Original	ML	Imputed	Original	ML	Imputed
Observations	249,739	587,449	618,375	72,523	152,992	161,559

Standard errors in parentheses

Clustered standard errors in parentheses.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2: Missingness in Instructor Gender

stu_gen1	-0.002	-0.001	-0.001	0.006	0.001	0.003
	(0.003)	(0.002)	(0.002)	(0.005)	(0.004)	(0.003)
Student is Asian	-0.004	0.001	0.001	0.005	-0.001	-0.002
	(0.004)	(0.004)	(0.003)	(0.008)	(0.005)	(0.004)
Student is White	-0.006	-0.001	-0.000	-0.001	-0.010	-0.009*
	(0.005)	(0.004)	(0.004)	(0.009)	(0.007)	(0.005)
Student is Black	-0.009	-0.002	0.005	-0.017	-0.020*	-0.009
	(0.008)	(0.007)	(0.006)	(0.015)	(0.011)	(0.009)
Student is Hispanic	-0.006	0.001	0.002	-0.004	0.002	-0.001
	(0.005)	(0.004)	(0.003)	(0.008)	(0.006)	(0.005)
Student_course_Term_GPA	-0.001	0.003	0.002	-0.003	-0.002	-0.001
	(0.002)	(0.002)	(0.002)	(0.004)	(0.003)	(0.002)
High_School_GPA	0.004	-0.001	0.003	-0.004	0.010	0.011**
	(0.005)	(0.004)	(0.004)	(0.009)	(0.007)	(0.005)
Student_current_age	-0.002	-0.002	-0.001	0.000	-0.003	-0.003
	(0.003)	(0.002)	(0.001)	(0.004)	(0.003)	(0.002)
response	0.000	0.000	0.000	0.000	0.000	0.000
	(.)	(.)	(.)	(.)	(.)	(.)
Scores for component 1	0.001	-0.003***	-0.002*	-0.001	-0.002	-0.000
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)
Sample	Freshman	Freshman	Freshman	Freshman, Fall	Freshman, Fall	Freshman, Fall
Gender Var	Original	ML	Imputed	Original	ML	Imputed
Observations	99,355	99,509	99,509	33,567	33,622	33,622
F-statistic	0.419	1.247	1.019	0.773	1.495	1.276
Prob ¿ F	0.926	0.261	0.422	0.642	0.144	0.245

All regressions include course fixed effects.

,

Standard errors are clustered at the instructor level.

Table 3: Balance Check. Dep Var: Instructor Gender

Student Gender(1=M)	0.004	0.001	0.002	-0.004	0.003	0.003
	(0.006)	(0.004)	(0.004)	(0.008)	(0.006)	(0.006)
Student is Asian	-0.005	0.001	-0.003	-0.002	0.002	0.001
	(0.006)	(0.004)	(0.004)	(0.008)	(0.006)	(0.006)
Student is White	-0.008	-0.007	-0.009*	-0.002	-0.014*	-0.012
	(0.008)	(0.005)	(0.005)	(0.011)	(0.009)	(0.009)
High_School_GPA	-0.007	0.003	0.002	-0.010	0.015	0.015
	(0.009)	(0.006)	(0.006)	(0.013)	(0.010)	(0.010)
Student_current_age	-0.003	-0.002	-0.003	-0.006	-0.005	-0.006
	(0.004)	(0.003)	(0.003)	(0.005)	(0.005)	(0.005)
FE	C, YxQ	C, YxQ	C, YxQ	C, YxQ	C, YxQ	C, YxQ
Sample	Freshman	Freshman	Freshman	Freshman, Fall	Freshman, Fall	Freshman, Fall
Gender Var	Original	ML	Imputed	Original	ML	Imputed
Observations	35,376	88,874	92,421	14,622	30,195	31,696
F-statistic	0.580	0.702	0.840	0.423	1.241	1.258
Prob ¿ F	0.716	0.622	0.521	0.833	0.287	0.279

All regressions include course fixed effects.

,

Standard errors are clustered at the instructor level.

Table 4: Regression Results of Instructor Student on PC

	All Sections						With More than 1 TA					
Female Student	0.076*	0.065	0.071***	0.062**	0.066***	0.057**	0.066*	0.057	0.070***	0.062**	0.064***	0.057**
	(0.040)	(0.040)	(0.025)	(0.025)	(0.025)	(0.025)	(0.040)	(0.041)	(0.025)	(0.025)	(0.025)	(0.025)
Female	-0.057	-0.054					-0.058	-0.057				
	(0.078)	(0.078)					(0.079)	(0.079)				
Female Student × Female	0.168***	0.161***					0.177***	0.170***				
	(0.056)	(0.056)					(0.057)	(0.057)				
Female			0.012	0.011					0.022	0.022		
			(0.050)	(0.050)					(0.051)	(0.050)		
Female Student × Female			0.132***	0.124***					0.136***	0.128***		
			(0.039)	(0.039)					(0.039)	(0.038)		
Female					-0.017	-0.016					-0.007	-0.007
					(0.049)	(0.048)					(0.049)	(0.049)
Female Student × Female					0.140***	0.133***					0.143***	0.136***
					(0.038)	(0.038)					(0.037)	(0.037)
$\hat{\beta}_2 + \hat{\beta}_3$	0.111	0.107	0.143***	0.136***	0.123**	0.117**	0.119	0.114	0.158***	0.150***	0.136***	0.129***
	(0.078)	(0.078)	(0.049)	(0.049)	(0.048)	(0.048)	(0.080)	(0.080)	(0.050)	(0.050)	(0.049)	(0.049)
Covariates	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Sample	Freshmnan	Freshmnan	Freshmnan	Freshmnan	Freshmnan	Freshmnan	Freshmnan	Freshmnan	Freshmnan	Freshmnan	Freshmnan	Freshmnan
Cluster												
Observations	35,410	35,376	88,998	88,874	92,549	92,421	33,650	33,621	84,605	84,488	87,970	87,850

All regressions include course and term fixed effects.

.

Standard errors are clustered at the discussion level.

Table 5: Regression Results of Instructor Student on PC

	All Sections						With More than 1 TA					
Female Student	0.046	0.033	0.073*	0.062	0.077*	0.066	0.054	0.040	0.075*	0.065	0.081*	0.071
	(0.064)	(0.063)	(0.044)	(0.044)	(0.043)	(0.043)	(0.065)	(0.065)	(0.044)	(0.044)	(0.043)	(0.043)
Female	-0.083	-0.085					-0.065	-0.068				
	(0.108)	(0.107)					(0.109)	(0.108)				
Female Student × Female	0.191**	0.185**					0.176**	0.172*				
	(0.087)	(0.087)					(0.089)	(0.089)				
Female			0.023	0.021					0.035	0.033		
			(0.074)	(0.074)					(0.074)	(0.074)		
Female Student × Female			0.113*	0.111*					0.110*	0.108*		
			(0.061)	(0.061)					(0.062)	(0.062)		
Female					-0.027	-0.028					-0.007	-0.009
					(0.074)	(0.074)					(0.074)	(0.074)
Female Student × Female					0.116*	0.113*					0.105*	0.102*
					(0.060)	(0.060)					(0.060)	(0.060)
$\hat{\beta}_2 + \hat{\beta}_3$	0.108	0.100	0.137*	0.132*	0.089	0.085	0.112	0.104	0.145**	0.140*	0.098	0.093
	(0.110)	(0.110)	(0.072)	(0.073)	(0.073)	(0.073)	(0.111)	(0.111)	(0.073)	(0.073)	(0.074)	(0.074)
Covariates	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Sample	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall
Cluster												
Observations	14,642	14,622	30,254	30,195	31,756	31,696	13,970	13,953	28,715	28,660	30,161	30,105

All regressions include course and term fixed effects.

.

Standard errors are clustered at the discussion level.

Table 6: Regression Results of Instructor Student on PC- No Covid

	All Sections						With More than 1 TA					
Female Student	0.046	0.033	0.073*	0.062	0.077*	0.066	0.054	0.040	0.075*	0.065	0.081*	0.071
	(0.064)	(0.063)	(0.044)	(0.044)	(0.043)	(0.043)	(0.065)	(0.065)	(0.044)	(0.044)	(0.043)	(0.043)
Female	-0.083	-0.085					-0.065	-0.068				
	(0.108)	(0.107)					(0.109)	(0.108)				
Female Student × Female	0.191**	0.185**					0.176**	0.172*				
	(0.087)	(0.087)					(0.089)	(0.089)				
Female			0.023	0.021					0.035	0.033		
			(0.074)	(0.074)					(0.074)	(0.074)		
Female Student × Female			0.113*	0.111*					0.110*	0.108*		
			(0.061)	(0.061)					(0.062)	(0.062)		
Female					-0.027	-0.028					-0.007	-0.009
					(0.074)	(0.074)					(0.074)	(0.074)
Female Student × Female					0.116*	0.113*					0.105*	0.102*
					(0.060)	(0.060)					(0.060)	(0.060)
$\hat{\beta}_2 + \hat{\beta}_3$	0.108	0.100	0.137*	0.132*	0.089	0.085	0.112	0.104	0.145**	0.140*	0.098	0.093
	(0.110)	(0.110)	(0.072)	(0.073)	(0.073)	(0.073)	(0.111)	(0.111)	(0.073)	(0.073)	(0.074)	(0.074)
Covariates	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Sample	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall	Freshmnan, Fall
Cluster												
Observations	14,642	14,622	30,254	30,195	31,756	31,696	13,970	13,953	28,715	28,660	30,161	30,105

All regressions include course and term fixed effects.

.

Standard errors are clustered at the discussion level.