Equilibrium Gender Discrimination and Disadvantage in Student Evaluations of Teaching

Sara Ayllón Lars J. Lefgren Richard W. Patterson
Olga B. Stoddard Nicolas Urdaneta

March 2025

Abstract

How should gender discrimination and systemic disadvantage be addressed when more discriminatory and less generous students systematically sort into certain fields, courses, and instructors' sections? In this paper, we estimate measures of gender bias and evaluation generosity at the student level by examining the gap between how a student rates male and female instructors, controlling for professor fixed effects. Accounting for measurement error, we find significant variation in gender bias and generosity across students. Furthermore, we uncover that bias varies systematically by gender and field of study and that patterns of sorting are sufficiently large to place female faculty at a substantive disadvantage in some fields and male faculty at a disadvantage in others. Finally, we document that sexist attitudes are predictive of gender-based sorting and propose Empirical Bayes inspired measures of student-level bias to correct for instructor-specific advantages and disadvantages caused by sorting.

1 Introduction

Performance evaluations are an integral part of many professional settings. They are consequential for promotions, pay, and hiring decisions and are widely used across all industries.[1] The subjective nature of performance evaluations, however, can lead to adverse labor market and career implications as evaluators' tendencies to be more or less generous, along with explicit or implicit biases towards an individual of a certain gender or race, may impact their assessment. Prior studies have documented gender bias in performance evaluations in many settings, including academic publishing (Card et al., 2019), recommendation letters (Eberhardt et al., 2023), government (Beaman et al., 2009), orchestra auditions (Goldin and Rouse, 2000), coding evaluations in the technology sector (Amer et al., 2024), online math forum (Bohren et al., 2019), and student evaluations of teaching (Boring, 2017; Mengel et al., 2019).

In academia, teaching evaluations are used to assess professors' performance and are key to their career progression. Notably, prior research investigating gender bias against female instructors using experimental and quasi-experimental methods has found conflicting evidence. While some studies document that female instructors on average receive significantly lower teaching evaluations than their male counterparts (Mengel et al., 2019; Keng, 2020; Boring et al., 2016; MacNell et al., 2015; Mitchell and Martin, 2018; Wagner et al., 2016; Ayllón, 2022), others find no evidence of bias against female instructors (Andersson et al., 2023; Binderkrantz et al., 2022; Acosta-Soto et al., 2022).

We propose two explanations for this puzzle, broadly defined as measurement and sorting. First, we argue that different empirical approaches used in existing research measure fundamentally different types of bias. What researchers have broadly referred to as gender bias fits into three categories that we define as (1) Fixed Input (FI) bias—raters scoring male and female ratees differently when ratees of both genders exhibit identical behavior,

[1] 74% of U.S.-based employees have formal performance reviews at least once a year (Wigert and Harter, 2017).

(2) Fixed Output (FO) bias—raters scoring male and female ratees differently when ratees of both genders generate identical output, and (3) Differential Valuation (DV) bias—raters scoring male and female ratees differently in an environment where performance is multi-dimensional and male and female ratees generate systematically different but equally socially valuable bundles of performance. Randomization of faculty gender identity in online settings (e.g., Andersson et al. (2023), Acosta-Soto et al. (2022) Baker et al. (2022), MacNell et al. (2015)) allows researchers to identify average levels of FI bias. In contrast, randomization of students to male and female instructors (e.g., Mengel et al. (2019) and Boring (2017))-can identify FO Bias if male and female instructors generate similar outputs or DV Bias if male and female instructor outputs have similar social values. Given that different empirical approaches measure different types of gender bias, there is no reason to expect different approaches to yield similar levels of bias.

The second explanation for the puzzle stems from students' ability to sort into institutions, majors, courses, and instructors on the basis of students' own bias. Consequently, the equilibrium effects of bias are likely to vary substantially across faculty based on the extent to which such sorting occurs. Indeed, one of the key insights of Becker (1971) is that the equilibrium level of discrimination an individual is exposed to may differ from the average level due to endogenous choices of market participants. In the context of student evaluations, the most biased students may sort to majors or classes in which they are exposed to comparatively few female faculty. We find evidence of sorting across multiple dimensions, including gender, field of study, and gender bias. For example, in our survey of U.S. undergraduates, we find that students who demonstrate more sexist attitudes toward women are less likely to take courses from female instructors–even controlling for their own gender and major. If student sorting is related to bias, then the average bias faculty face may be smaller than the average levels of discriminatory preferences would suggest.

Finally, in addition to bias, student sorting may also lead to substantial variation in the overall rater generosity that faculty face, leading to variation in subjective evaluations

unrelated to instructor effort or skill. Hence, even when using the same definition of bias, sorting patterns may lead to considerable inconsistencies in bias estimates and vary by setting (eg. men vs. women, fields, types of courses). One potential way to reconcile the mixed results of prior studies is to account for sorting of students to institutions, majors, courses, and instructors.

Once one identifies bias and persistent generosity at the student level, it becomes possible to examine patterns of sorting and characterize situations in which specific faculty are plausibly disadvantaged. We develop these insights by adapting the empirical framework of Abowd et al. (1999), in which we regress student-level ratings on professor and student fixed effects. We further modify this framework to allow students to have different levels of generosity towards male and female instructors.

This empirical framework provides several important contributions to the existing literature. First, we are able to characterize the distribution of gender bias across students, accounting for measurement error, and show that the average level of bias found in prior work is not constant across students. Rather, it varies both idiosyncratically and systematically by student gender. Specifically, female students exhibit significant relative favoritism towards female faculty. Second, we examine the sorting of students both within and across fields of study. We show that female students sort to fields with more female faculty. Consequently, the average female professor teaches students who are predicted to be more sympathetic than the student body as a whole. Within field, we again see female students sort disproportionately to female faculty. Third, we construct measures of gender-specific generosity at the student level, which are approximately forecast unbiased. Using these tools allows policymakers both to identify specific contexts in which female and male faculty face a particularly severe disadvantage and make corrections as needed.

In addition to contributing to an important literature on gender bias in performance evaluations, our findings suggest that evaluation research has not paid enough attention to persistent differences in evaluator generosity. We show that students' predicted generosity

strongly impacts female instructors' evaluation outcomes. Specifically, female faculty exposed to the bottom-quartile of student gender-specific generosity are 70% more likely to receive the bottom-quartile of overall student rankings. In contrast, students with top-quartile draws of predicted generosity are nearly twice as likely as those with bottom-quartile draws to receive top-quartile overall student evaluations. Hence, adjustments for gender-specific generosity can be large and important.

2 Defining Gender Bias

What do researchers mean when they discuss gender bias? While the term is broadly used, researchers often have different implicit definitions in mind. In this section, we identify three types of bias and conditions under which each type of bias would be identified. It should be noted that the definitions of bias are neither mutually exclusive nor nested.

2.1 Fixed Input (FI) Bias

We define student i's Fixed Input (FI) bias with the following equation:

$$FIBias_i = E[R_{it}|I_t = \bar{I}, g_t = male] - E[R_{it}|I_t = \bar{I}, g_t = female] \tag{1}$$

where R_{it} is a rating individual i gives to an instructor t, I_t is a vector of instructor t's inputs, $\bar{I}$ is fixed level of instructor inputs, and g_t is a variable indicating the gender of an instructor t. If the inputs of faculty are held constant across faculty gender and an individual i gives different ratings to male and female instructors *despite* identical inputs across male and female faculty, then this individual has FI bias.

Studies such as Andersson et al. (2023), Acosta-Soto et al. (2022) and MacNell et al. (2015) that randomly assign the online gender identity of instructors offer a clear test of FI bias. The evidence on fixed input bias is mixed. Unless researchers directly observe instructor inputs or make strong assumptions about the relative inputs of male and female instructors,

identifying FI Bias requires the ability to hide the actual gender of an instructor, necessarily limiting the scope of instruction that can be assessed. FI is best assessed in instruction that is online, asynchronous, and text-based while virtually impossible assess in in-person, synchronous, and interactive environments. Additionally, because FI bias only applies to situations where inputs are fixed, the relevance of FI bias is limited if female and male instructors, on average, have significantly different inputs in their teaching production.

2.2 Fixed Output (FO) Bias

To the extent that female and male instructors, on average, have different inputs, we can expand our definition of bias to allow different inputs and, instead, hold output constant. We define i's Fixed Output Bias as:

$$FOBias_i = E[R_{it}|O_t = \bar{O}, g_t = male] - E[O_{it}|O_t = \bar{O}, g_t = female] \tag{2}$$

where R_{it} is a rating individual i gives to an instructor t, O_t is instructor t's output, $\bar{I}$ is fixed level of instructor output, and g_t is a variable indicating the gender of an instructor t. While conceptually attractive, directly measuring FO Bias is difficult precisely because it requires knowing or making strong assumptions about instructor output. If we have reliable measures of instructor output, then student evaluations have trivial value. If reliable measures of output are not available but researchers are willing to make assumptions about the relative output of male and female instructors (e.g., conditional on observable characteristics, female and male instructors produce the same output), then studies that randomize students to female and male instructors (e.g. Boring (2017); Mengel et al. (2019)) can identify the presence of FO Bias by examining the differences in male and female instructor ratings.

2.3 Differential Valuation (DV) Bias

In reality, instructor output is multi-dimensional (e.g. instructors produce test performance, writing improvement, persistence, job-readiness, etc.) and it is likely that female and male instructors provide different baskets of goods to students. The greater the divergence in inputs and outputs of male and female instructors, the less relevance measures of FI Bias and FO Bias will have. Differential valuation (DV) Bias provides a framework for assessing bias when the male and female outputs diverge as long as researchers are willing to make a value judgment about the outputs that male and female instructors generate. For example, one reasonable value judgment is that, on average, male and female instructors generate equally socially valuable outputs. If a student rates female and male instructors differently when female and male instructors generate the same social value, then the student exhibits DV Bias. Specifically, We define student i's DV bias as:

$$DVBias_i = E[R_{it}|O_t\Gamma = \overline{O\Gamma}, g_t = male] - E[R_{it}|O_t\Gamma = \overline{O\Gamma}, g_t = female] \tag{3}$$

where R_{it} is a rating individual i gives to an instructor t, I_t is O_t is a vector instructor t's outputs, Γ is a vector of coefficients that capture the social welfare value of each output $\overline{O\Gamma}$ is fixed value of instructor outputs, and g_t is a variable indicating the gender of an instructor t. As long as male and female instructors' outputs are equally socially valuable, then DV bias is measurable and arises when students rate male and female instructors in ways that systematically deviate from the welfare values of their male and female instructors' outputs. Once we make an assumption about the relative social value of female and male instructors (e.g. conditional on observable characteristics, female and male instructors produce the same social value), randomization of students to male and female instructors allows the identifies the combination of FI, FO, and DV bias. Unless one is able to condition on instructor inputs and outputs, isolating DV bias from FI and FO bias is not feasible even with randomization.

3 Survey Evidence of Student Sorting

While gender bias can be identified under random assignment, generally students sort non-randomly across classes and instructors. This suggests that the bias faced by any particular instructor may be very different than the average bias in the population. Consequently, we begin our empirical analysis by exploring patterns of student sorting. To provide evidence on whether students sort to faculty based on sex and gender bias, we surveyed 359 American undergraduate students on an online platform, Prolific, about their academic schedules and majors. Specifically, we ask them to list the four classes they had taken most recently and indicate the demographic characteristics of the instructors. We also elicit their gender attitudes by asking four questions from the standard Glick and Fiske (1996) ambivalent sexism scale.[2]

As reported in Figure A.1, we reject the hypothesis that sexist attitudes are uniformly distributed across fields ($p<0.01$) with students in Arts and Communications majors exhibiting exceptionally low rates of sexism and students in Business and Economics majors exhibiting exceptionally high rates of sexism. Next, we show that female students sort toward female faculty. Specifically, as reported in column (1) of Table 1, female students have a 13.5 percentage point greater female faculty share than male students in our survey. In column (2) we find that while major choice explains approximately 20% of the male-female gap in female faculty share, there is still substantial gender-based sorting within degree field. In columns (3) and (4) of Table 1 we explore whether gender-based sorting is occurring by examining the relationship between sexist attitudes and female faculty share. We find that a standard deviation increase in our measure of student sexism corresponds to a 4 percentage point lower female faculty share, regardless of whether we account for student majors ($p<0.01$). Finally, in columns (5) and (6) we examine whether apparent sexism-based sort-

[2]These questions ask students how much they agree with the following statements: (1) No matter how accomplished he is, a man is not truly complete until he has the love of a woman; (2) Many women are actually seeking special favors, such as hiring policies that favor them over men, under the guise of asking for equality; (3) Women are too easily offended; and, (4) Many women have a quality of purity that few men possess. We describe the survey in more detail and outline the protocol in Appendix C.

ing can simply be explained by gender-based sorting, as female students have significantly less sexist attitudes than their male peers. Our findings suggest that student gender can only explain about a quarter of sexist-based sorting: after accounting for student gender, a standard-deviation increase in sexist attitudes predicts a 3 percentage point lower female faculty share ($p<0.05$).

The fact that students with sexist attitudes sort away from female faculty implies that, on average, female faculty are shielded from some sexism and encounter less sexism than they would if instructors were randomly assigned. It also suggests, however, that the sexism individual female faculty members depends on their institution, field, and any institutional factors that affect how students sort to classes and faculty. While this effectively documents both heterogeneity in sexist attitudes and non-random sorting of students to instructors, it is important to develop a method to measure how this impacts specific faculty in actual academic settings, which we do in this paper.

4 Data

Our data comes from the universe of student evaluations of teachers at the University of Girona (Spain) from Fall 2015 to Fall 2022.[3] At this university, students are asked to complete an anonymous online teaching evaluation questionnaire at the end of each semester. Completing the questionnaire is not mandatory but students are reminded about the importance of the evaluations in their classes and by email. The questionnaire is administered in all courses and for nearly all instructors.[4]

In total, we observe 328,429 student evaluations. We restrict our analysis to evaluations that rate an instructor's overall performance and that can be linked to an instructor's characteristics. We also drop evaluations from students who do not have a stated major or who

[3]The University of Girona is a public university in Spain. It enrolls approximately 16,000 students per academic year and employs 1,200 academic staff. It has 10 Colleges and 24 Departments that teach individual degree programmes at Bachelor's, Master's, and Doctoral levels.

[4]Questionnaires are administered for instructors who have taught at least 1.5 European Credit Transfer System (ECTS) credits (0.75 U.S. academic credits) in a course.

are enrolled in a small specialty program. This leaves us with a sample of 263,460 evaluation records from 15,862 students, 27,381 course sections, and 2,902 instructors.

Our main outcome variable is the response to the statement 'I evaluate this teacher's overall performance as positive.' Students can answer on a scale of 1 to 5, ranging from 'strong disagreement' to 'strong agreement'. Panel A of Figure A.2 in the Appendix shows that high ratings are much more common than low ratings, with 43% of responses strongly agreeing with the statement about positive performance and only 6% strongly disagreeing with this statement. Panel B Shows the distributions of average ratings for each student and course, with the median student-average rating of 4.1 and the median course-average rating of 4.2.

Table A.1 provides summary statistics for our estimation sample. Similar to many other universities, the majority of students in our sample are female (59%). About 42% of student sections are taught by female instructors. This differs substantially by the gender of the student. 47% of classes taken by female students are taught by female faculty while the corresponding rate for male students is only 35%. This reflects both sorting within majors as well as sorting across majors. For example, female students and faculty are underrepresented in fields such as Business, Engineering and Economics and are overrepresented in Education, Medicine, and Social Work. We also observe that female students award slightly higher student evaluations on average than their male peers. Differences in the gender composition of students by instructor gender and field along with differences in how female and male students rate faculty suggest that student sorting could be an important factor in instructor ratings.

5 Examining Average Differences in Ratings between Male and Female Professors

Before developing our framework for estimating student-level generosity and bias, we first document average differences in ratings by professor gender. To do so, we normalize our primary measure of instructor performance to have a mean of zero and a standard deviation of 1. We regress this normalized outcome measure on instructor gender in column (1) of Table 2 and find that women receive slightly higher ratings than men, although the difference is not statistically significant.[5] Thus, *after* student and faculty sort into fields and courses, female instructors at the University of Girona receive higher ratings than male instructors on average. What happens as we start to 'undo' student and faculty sorting? In columns 2-5, we successively add controls for faculty characteristics (age, tenure, contract status), field and course characteristics, student characteristics (gender, age, degree program, and whether they have taken the course before), and final course grades. Each successive set of controls reduces the female instructor coefficient, moving from female instructors receiving 0.015 SD higher ratings than male instructors in column (1) to 0.046 SD *lower* ratings in column (5), a difference that is both economically and statistically significant ($p<0.05$).

Many experimental and quasi-experimental studies have documented bias against women at the student level (e.g. Boring et al., 2016; MacNell et al., 2015; Mengel et al., 2019; Mitchell and Martin, 2018; Wagner et al., 2016; Keng, 2020; Fan et al., 2019). We find comparable results only after accounting for faculty, student, and course characteristics. One explanation for this finding is that female students who view female faculty most favorably are disproportionately likely to sort into classes with female instructors. It may also be that female faculty sort into fields or are assigned to teach courses that draw more generous students in general. If these types of student and faculty sorting can explain our Table 2 results, a female instructor who teaches a class in a male-dominated field that draws relatively ungenerous

[5] These estimates are weighted at the instructor level to make them comparable to the estimates we report later, though the results are qualitatively similar when we weight either at the response or course level.

students would still be subject to substantial gender bias and disadvantage. Applying Becker (1971)'s theory of equilibrium discrimination, small differences researchers frequently observe in average ratings may mask significant variability in the bias and disadvantage individual instructors face.

6 Applying the AKM Model to Student Evaluations

In this section, we present a statistical framework for measuring ratings generosity and gender bias at the student level. In this framework, generosity is defined as the empirical propensity, *relative to other students*, to give high evaluations to a fixed set of faculty. Gender bias is then defined as the difference between a student's generosity toward male faculty and their generosity toward female faculty. This metric captures whether a student, relative to their peers, is more generous to male faculty than to female faculty. Importantly, this measure of relative bias across students is identifiable, even if the average absolute levels of performance for male and female instructors are not. Moreover, bias may reflect either an affinity for faculty based on gender or a preference for instructor behaviors that, on average, differ across male and female faculty. If one student is more biased than another student according to any of the definitions we propose, this will be reflected in the empirical measure of bias we estimate.

Formally, we consider the following regression model:

$$R_{tci} = Z_c\Pi + \theta_t + \phi_i + \nu_{tci} \tag{4}$$

In this model, R_{tci} is the rating given to teacher t in classroom c by student i. Z_c is a vector of class characteristics such as the semester in which the class is taken. θ_t captures the fixed observed and unobserved characteristics of teacher t, including the average effectiveness of the instructor. ϕ_i captures the rating generosity of the student. ν_{tci} is the idiosyncratic component of the rating, which captures the quality of match between student and instructor.

This empirical specification is an application of the AKM methodology developed in Abowd et al. (1999).[6]

Consistent identification of instructor (θ_t) and student (ϕ_i) effects requires two conditions. First, estimation must be performed on a connected set of students and instructors. Because students at the University of Girona infrequently take courses outside of their majors, we estimate separate models for 20 major types that include students taking many courses in common. In doing so, we give up on the possibility of comparing average generosity and effectiveness across substantially different fields but benefit from estimating the parameters of our model within a richly connected set. Estimation within field also helps alleviate concerns regarding endogenous sorting of students to faculty on the basis of idiosyncratic interest in course content.

The second key assumption underlying our model is that students cannot sort based on the idiosyncratic match quality of the student and instructor. This holds, by construction, in settings with random assignment of students to instructors. However, in many settings this assumption will be violated as students sort to instructors who match their preferred teaching styles or to same-gender instructors. To explicitly deal with sorting on gender match, we estimate separate statistical models for male and female instructors, which yields student-specific generosity measures for male and female instructors. This regression model is given by the following:

$$R_{tgci} = Z_c\Pi_g + \theta_{tg} + \phi_{gi} + \nu_{tgci} \tag{5}$$

The terms of this equation are similar to those in Equation 4 besides the fact that they are indexed by the gender of the faculty. Note that the average of θ_{tg} is not separately identified from the average of ϕ_{gi} because gender does not (often) vary within an instructor. Consequently, our analysis can identify relative ratings generosity towards male faculty versus female faculty but not the average level of bias without additional assumptions. Our measure

[6]In the case of the AKM model, wages are decomposed into firm and worker effects.

of bias is the difference in student generosity towards male and female faculty or $b_i = \phi_{mi} - \phi_{fi}$.

This enriched model explicitly allows for selection of students to instructors on the basis of faculty gender. However, there may still be sorting across instructors within gender on the basis of the idiosyncratic match quality between faculty and students. To address this concern, we examine the correlation of student generosity measures in a subset of required courses (over which students have little discretion) and all courses a student takes. Adjusting for estimation error, correlation between student generosity measures calculated on the sample of required courses and non-required courses is indistinguishable from 1, suggesting that the sorting of students to classes based on idiosyncratic preferences is not a significant source of bias.

One additional potential selection issue in our empirical example is that students are not required to complete evaluations. For example, students may be more likely to submit evaluations for instructors that they view as particularly effective than to submit evaluations for faculty they view as ineffective. In this case, the student's estimated generosity measure would be positive. However, the student's latent generosity, if compelled to evaluate all instructors, would be lower. We don't believe this to be a substantial issue, however. Consider a simulation in which 50 percent of students *only* are twice as likely to report when their experience is positive relative to when it's negative. In this case, the correlation between a student's observed and latent measures of generosity is 0.94.

In the subsections below, we discuss how we use our empirical measures to characterize the distribution and predictors of generosity and bias. We also explore patterns of sorting across sections. Finally, we present a method for adjusting course ratings to take into account the generosity and bias of students in the courses.

6.1 Estimating the Variance of Generosity and Bias

Due to the fact that each student interacts with a finite number of faculty, our measures of generosity and bias are necessarily quite noisy. Consequently, the variance of these raw measures greatly overstates the actual variability of generosity and bias. To overcome this challenge, we stratify by instructor gender and randomly split instructors into two subsamples. We then estimate Equations 4 and 5 for each subsample. This yields two noisy measures of each student's measures of generosity and bias. As long as the estimation error in each of the estimates is independent, the following equality holds.

$$\sigma^2_\phi = cov(\hat{\phi}^1, \hat{\phi}^2) \tag{6}$$

where $\hat{\phi}^1$ and $\hat{\phi}^2$ represent the estimates of student generosity from the first and second splits of the data.[7] This allows us to estimate variance of latent generosity via a method of moments estimator in which we simply calculate the sample covariance between the estimated measures of generosity and bias from the two splits of data. This yields the estimate $\hat{\sigma^2}_\phi$.

The first row of in Panel A of Table 3 shows the standard deviation of empirical estimates of generosity and bias. These measures are the sum of latent generosity and bias along with estimation error. In the second row of results, we show estimates of the latent measures of generosity and bias as calculated using equation 6. We see that the standard deviation of overall latent generosity is 0.340, implying that a student who is one standard deviation higher in the generosity measure would give, on average, ratings that were about 0.340 standard deviations higher to a given professor than the average student. The measures are quite similar when looking at generosity towards male and female faculty. The standard deviation of bias is 0.207. Relative to the average student, one with a bias measure one standard deviation higher would tend to give male professors 0.207 standard deviation higher ratings than female professors. To put this into perspective, in our primary setting evaluations are

[7] $\hat{\phi}_i$ is an empirical analog of ϕ_i from Equation 4.

given on a five-point scale. Our estimates suggest that a student with a bias measure one standard deviation higher than average might be 16 percentage points more likely to drop the evaluation of a female professor by 1 point than of a male professor.

6.2 Predicting Student Generosity and Bias

It is not only interesting to document the existence of variation in student generosity and bias but also useful to consider whether student characteristics are predictive of generosity and bias. To do so, we estimate the following regression equation:

$$\hat{\phi}_i = X_i\beta + \epsilon_i \tag{7}$$

In this equation, X_i represents a vector of observable student characteristics including student gender and age. The coefficient β indicates which observable factors are predictive of overall student generosity. We estimate analogous regressions when calculating which factors are predictive of gender-specific generosity as well as bias.[8] In Panel B of Table 3 we estimate that female students give 0.046 standard deviations more generous ratings than male students on average ($p<0.01$). Female students are similarly generous to male faculty as male students but are 0.087 standard deviations more generous to female faculty than male students ($p<0.01$). Consequently, female students are on average 0.075 standard deviations less biased towards male faculty than male students ($p< 0.01$). In Panel B of Table 3 we also find that older students are more generous in general, but particularly more generous toward female faculty. For every year older a student is she or he is 0.011 standard deviations more generous toward male faculty, 0.015 standard deviations more generous toward female faculty and 0.004 standard deviations less biased towards male faculty (all estimates significant at $p< 0.01$).

[8]Note that we do not observe gender-specific generosity measures for those students who have not evaluated a class by a professor of the relevant gender. We also do not observe bias measures for students who have not taken courses from both male and female faculty. Consequently, the predictions are conditional upon having taken at least one class from a faculty member of the relevant gender.

6.3 Are Estimates of Generosity and Bias Predictive Out of Sample?

Our estimates suggest substantial variability of generosity and bias within our sample. These estimates are particularly useful if they are predictive out of sample, allowing researchers and practitioners to identify what faculty are subject to significantly biased or ungenerous students. However, the predictability of such estimates is substantially reduced by estimation error. Motivated by the empirical Bayes shrinkage estimates (Morris, 1983), we overcome this limitation by implementing a procedure to isolate our predictions of generosity and bias from estimation error that we outline in Section B.

In our approach, we construct estimation-error adjusted predictions of generosity $\tilde{\phi}_i^C$ and bias $\tilde{bias}_i^C$ using evaluations from the 2015-2020 school years. Then, to evaluate whether these estimates are predictive out-of-sample, we estimate the following equations:

$$\hat{\phi}_i^D = \alpha_0 + \alpha_1 \tilde{\phi}_i^C + e_i \tag{8}$$

and

$$\hat{bias}_i^D = \beta_0 + \beta_1 \tilde{bias}_i^C + e_i \tag{9}$$

Where $\hat{\phi}_i^D$ and $\hat{bias}_i^D$ are estimates of individual i's generosity and bias in 2021 respectively. If our estimates of $\tilde{\phi}_i^C$ and $\tilde{bias}_i^C$, are predictive out-of-sample we expect both α_1 and β_1 to be positive and statistically significant. Furthermore, if our estimates of $\tilde{\phi}_i^C$ and $\tilde{bias}_i^C$ are forecast-unbiased then we expect α_1 and β_1 to be insignificantly different from 1.

We examine our estimates of Equations 8 and 9 in Table 4. In column (1), we find not only that our predictions of overall student-level generosity are predictive out-of-sample ($p<0.01$), but they are forecast-unbiased: our estimate of α_1 is 0.988 and not statistically distinguishable from 1 ($p=0.78$). When we estimate faculty gender-specific versions of Equation 8, in columns (2) and (3) of Table 4 we find that gender-specific predictions of generosity ($\tilde{\phi}_{im}^C$ and

$\tilde{\phi}_{if}^{C}$) are similarly predictive out-of-sample and forecast-unbiased.[9] While our predictions of generosity appear to be both predictive out-of-sample and forecast-unbiased, our predictions of bias are only predictive out-of-sample and not forecast-unbiased. Specifically, in column (4) our estimate of β_1 is 0.493 and, while significantly different from zero (p<0.01), it is also significantly different from 1 (p<0.01). The primary challenge to obtaining forecast-unbiased predictors of gender bias is that a student's predicted generosity toward male professors ($\tilde{\phi}_{im}^{C}$) is highly correlated with a student's generosity toward female professors ($\tilde{\phi}_{if}^{C}$). If we restrict our test to only individuals with student evaluations from the prior year, our measures of student generosity and bias perform even better. The coefficient on our predictions of student generosity range between 0.99 and 1.02 and are all statistically insignificantly different from 0. The coefficient on our prediction of bias is 0.62 and also statistically insignificantly different from 1 at the 5 percent level. Taken together, this evidence suggests that student generosity is both stable over time and predictive of future behavior.

It should be noted that in order for our measures of gender-specific generosity to be forecast unbiased, it must be the case that the ratings regime is stable across periods. When the ratings regime changes, this will tend to change (usually reduce) the correlation between student generosity measures estimated before and after the shift. Depending on when the regime shift occurs, this can lead to the coefficient on predicted generosity being significantly above or below 1.

7 The Bias Experienced by Female Faculty

7.1 Measuring Average Bias

Having developed a structure for estimating student generosity and bias, it is helpful to think about how to determine the extent to which female faculty are affected, on average,

[9] Our estimate of $\tilde{\phi}_{if}^{C}$ of 0.905 is marginally different from zero. However, when we estimate the predictive power of ratings-level generosity forecasts in columns (5)-(7), we find that they are predictive and forecast-unbiased.

by student bias. This exercise is complicated, however, by the fact that we cannot observe the underlying effectiveness of male and female faculty. Consequently, if ratings are higher for one group than another, one cannot rule out that observed differences in ratings reflect underlying differences in average effectiveness. If one is willing to assume that men and women are equally effective (i.e. make FO bias assumptions) or provide equally valuable outputs (i.e. make DV bias assumptions), one can test for bias by examining whether female instructors receive lower ratings, holding fixed the composition of students who rate them. Furthermore, assuming equal effectiveness is supported by some existing evidence. For example, at the United States Air Force Academy, students are randomly assigned to introductory math and science instructors and all instructors teach a common syllabus, administer common exams, and pool grading tasks with other instructors. In this setting Carrell et al. (2010) find that while a random assignment to a female instructor leads female students to do somewhat better and male students to do somewhat worse, these differences are offsetting and there is no difference in overall performance by instructor gender.

Recall that the AKM model allows for sorting on the basis of persistent student generosity and teacher effectiveness. Under the assumption of no sorting on the *idiosyncratic* match quality between students and instructors,[10] including matching on gender-bias, we can leverage Equation 4 to determine instructor effectiveness, controlling for the ratings generosity of the students in their classes. In this case, the average rating of an instructor can be decomposed in the following way (for simplicity, we abstract from covariates and estimation error):

$$\bar{R}_t = \theta_t + \bar{\phi}_t \tag{10}$$

Note that $\bar{R}_t = \frac{\sum_{i=1}^{n} R_{ti} 1_{ti}}{n_t}$, where R_{ti} is the rating by student i to instructor t, 1_{ti} is an indicator variable that takes a value of 1 if student i rated teacher t, and n_t is the total

[10]By construction, the empirical residuals will be uncorrelated to the estimated teacher and student fixed effects.

number of students taught by teacher t. $\bar{\phi}_t = \frac{\sum_{i=1}^{n} \phi_i 1_{ti}}{n_t}$ is the average generosity measure of students taught by teacher t. This equation shows that, if the assumptions of the baseline AKM model hold, gender differences in ratings reflect both the tendency of female faculty to receive higher or lower ratings from a given set of students as well as the composition of students they teach. We can examine this decomposition by estimating regressions in which the dependent variable is either the estimated teacher fixed effect or the average student generosity measure of students taught by the teacher. The independent variable of interest is a female professor indicator variable.

In columns (1) and (2) of Table 5, we examine the within-degree differences in faculty ratings by instructor characteristics. In column (1) we see that female faculty receive 0.010 SD lower evaluations than male faculty on average, although the difference is not statistically significant. However, when we account for faculty age and permanence status, we find that female instructors receive 0.036 SD lower ratings ($p<0.10$). In columns (3) and (4) we regress faculty fixed effects onto faculty gender to isolate the component of differential ratings that comes from gender bias. In column (3), we observe that female faculty have 0.018 standard deviations lower ratings than male faculty when teaching the same students, although the difference is not statistically significant. When we control for faculty characteristics, gender bias grows to 0.044 standard deviations and becomes statistically significant ($p<0.05$).

7.2 Sorting on Generosity and Gender Bias

Do more generous students sort to certain types of faculty? In columns (5) and (6) of Table 5, whether more generous students sort into courses based on instructor sex, age, and status. Specifically, we show results from a faculty-level regression in which the dependent variable is the average student effect taught by the professor. While there does not appear to be much sorting on instructor age or status, we do find that more generous students sort into courses taught by female faculty: female faculty teach students who award ratings 0.013 standard deviations higher than those taught by male faculty. This sorting is beneficial, on average,

to female faculty and helps explain why average teacher ratings are similar between male and female faculty even in the presence of potential bias. Notably, our application of the AKM model only examines sorting within major. In Appendix Figure A.3 we find significant variation in the average ratings across majors, with instructors in the lowest-rated majors (e.g. architecture and economics) receiving approximately 0.5 SD lower ratings than the highest-rated majors (e.g. medicine and philosophy). While it is possible that differences in instructional quality explain these gaps, it is likely that student sorting plays an important role.

In addition to sorting on generosity, our survey evidence suggests students may select classes on the basis of preference over the gender of their instructor. We now examine the extent to which this occurs in our sample at the University of Girona. To answer this question, we leverage estimates from Equation 5. We test whether sorting occurs by regressing average bias measure of students taught by a particular faculty member on whether the professor is female. Columns (7) and (8) of Table 5 report these coefficients. Note that a negative coefficient on the female faculty indicator variable suggests that women have students who are more favorable to them, on average, than to male faculty. Consistent with this explanation, we see that female faculty have students who are less biased towards men than average. The coefficients are statistically insignificant, however.

Our failure to detect statistically significant sorting based on gender bias may reflect that little such sorting occurs, that our measures of bias are sufficiently noisy that we lack the statistical power to detect the sorting that occurs, or that our AKM approach only allows us to examine sorting within degrees. We do, however, observe strong sorting of female students to female faculty. In Figure 1, we show the relationship between the fraction of students who are female and the fraction of faculty who are female across degrees. Unsurprisingly, there is a strong positive relationship. Indeed, the estimate of the regression line yields a coefficient greater than 1. In Figure A.3 The sorting also occurs within degree field, where students have additional flexibility to select courses specifically on the basis of faculty gender. In columns

(9) and (10) of Table 5 we regress the fraction of students taught by a female professor on whether the professor is female. We see that, within degree fields, female professors are evaluated 3.5 percentage points more by female students than male professors in the same field. As a consequence of this sorting across and within fields, a female faculty member's fraction of female students is 0.71 compared to 0.57 for the overall sample of students.[11] Given that female students' measure of bias is 0.075 standard deviations lower than male students, the gender-based sorting of students to faculty suggests gains to female (and male) faculty of approximately 0.011 standard deviations in ratings. While this may seem small, it is on the same order as the unconditional gender gap in average ratings that occurs within our sample.

7.3 Identifying Settings in Which Female Faculty Are at a Disadvantage

Our findings from the University of Girona and our survey of U.S. college students together lead us to several conclusions. First, students vary substantially both in terms of how generous they are to male and female faculty as well as the degree of bias they exhibit towards female faculty. Second, sorting to female faculty based on student sex and student gender-bias likely attenuate the degree to which female faculty suffer from student bias, on average. However, there are likely specific settings in which female faculty are at a substantial disadvantage relative to their male counterparts. For example, female faculty in Business and Economics fields are likely to face substantially more gender-biased students than faculty in Arts and Communications fields and, as a result, receive significantly worse student ratings.

Furthermore, the disadvantage female faculty face varies predictably and substantially across instructors within a field as well. In our University of Girona sample, we can use our predictions of student-specific generosity toward female instructors within a field from

[11] From a student's perspective, female students in our sample take 47.2% of their courses from female instructors whereas male students only take 34.3% of their courses from female instructors.

Section 6.2 to examine the degree to which female faculty are disadvantaged by being exposed to students who do not give generous ratings to female faculty. In Panel A of Figure 2, we plot the actual cumulative density functions (CDFs) of average faculty ratings ($\bar{R}_t$) for female faculty who are either exposed to bottom- or top-quartile draws of predicted student generosity toward female instructors ($\bar{\phi}_{tf}$). This plot highlights two important points. First, our procedure generates accurate out-of-sample predictions of student-level generosity. The faculty with top-quartile draws of predicted student-generosity have higher actual ratings than faculty with bottom-quartile draws of predicted student generosity at every point of the distribution. Second, female faculty who draw students who are predicted to be less generous are at a significant disadvantage relative to female faculty who draw students who are predicted to be more generous. Relative to female faculty with top quartile draws of gender-specific predicted generosity, female faculty with bottom-quartile draws are 70% more likely to be in the bottom quartile of overall student evaluations of teachers (29.4% vs. 17.3%). In contrast, female faculty with top-quartile draws of gender-specific predicted generosity are nearly twice as likely as those with bottom-quartile draws to receive top-quartile overall student evaluations (35.6% vs. 17.9%).

If predictable variation in student generosity toward female instructors is not accounted for, a significant portion of variation in student evaluations will be misattributed to teacher quality, harming female faculty with 'bad' student draws and helping female faculty with 'good' student draws. Fortunately, we can construct ratings that are adjusted for gender-specific generosity, which implicitly corrects for bias ($\bar{R}^*_{tf} = \bar{R}_{tf} - \bar{\phi}_{tf}$). Because differences in the average gender-specific generosity of students is not separately identified from differences in average effectiveness of male and female instructors, we again lean on either an assumption of average equal effectiveness or value between genders. The use of gender-specific generosity measures allows for sorting of students to faculty on the basis idiosyncratic preferences over faculty gender. This adjustment effectively eliminates disadvantages caused by student

composition.[12] In Panel B of Figure 2, we plot the CDFs of generosity- and bias-adjusted faculty ratings ($\bar{R}_t^*$) for those with bottom- and top-quartile draws of predicted gender-specific generosity and find that the CDFs of the two groups are indistinguishable. Thus, our approach gives policy-makers a tool to appropriately adjust ratings for differences in the composition of students each faculty member faces.

We note that the performance of our approach is not guaranteed and depends on factors such as the stability of rankings over time. We apply our approach for earlier years (2018-2020) in our data in Table A.2 and Figure A.4. Table A.2 indicates that for 2018 and 2020 our approach does well at forecasting generosity but poorly at forecasting bias. For 2019 our approach does less well at forecasting generosity but does well at forecasting bias. Figure A.4 shows that even when our forecasts are imperfect, using our correction method serves to close the gap between female faculty with high female-specific generosity and low female-specific generosity.

8 Discussion

In this paper we estimate measures of gender bias and evaluation generosity at the student level by examining the gap between how a student rates male and female instructors, controlling for professor fixed effects. We propose two explanations for the conflicting results found in related prior work. First, we argue that different empirical approaches used in existing research measure fundamentally different types of bias. We define three types of bias - Fixed Input, Fixed Output, and Differential Valuation, to help reconcile existing findings. Second, we the equilibrium effects of bias are likely to vary substantially across faculty based on students' ability to sort into institutions, majors, courses, and instructors on the basis of their own bias.

We document substantial predictive variability in student generosity and gender bias

[12] Gender-specific generosity is measured at the field of major (connected set) level and is, therefore, common among all faculty within the same field. It is calculated to equalize adjusted ratings across male and female faculty within the field.

in evaluations of teaching. Consistent with homophily between students and faculty, we find that female students exhibit significantly less bias against female faculty than do male students. Most of the variability in gender bias is idiosyncratic at the student-level. Point estimates suggest students biased against female faculty sort away from such faculty, though the bias estimates are sufficiently noisy that these estimates lack statistical power. We show, however, that female students, who are less biased against female faculty on average, strongly sort to female faculty both across and within fields. We replicate this finding in a separate sample of U.S. college students and show that female students have approximately 13 percentage points greater female faculty share than male students. Collectively, these results suggest that the bias experienced by female faculty is moderated by the endogenous sorting of students across fields and classes.

We find considerable variability in the disadvantage faced by female faculty across and within fields. Women faculty exposed to a class with students with low female-specific generosity perform substantially worse on average than female faculty with more sympathetic students. Specifically, relative to female faculty with top-quartile draws of student gender-specific predicted generosity, female faculty with bottom-quartile draws are 70% more likely to receive bottom-quartile student evaluations while female faculty with top-quartile draws of gender-specific generosity are nearly twice as likely as those with bottom-quartile draws to receive top-quartile overall student evaluations.

To investigate the extent to which gender attitudes contribute to the observed gender-based sorting we conduct an online survey of U.S. college students. We find striking variability in gender attitudes across fields, with students in Business and Economics exhibiting 0.95 standard deviation higher levels of sexism than students in Arts and Communications majors. While major choice explains approximately 20% of the male-female student gap in female faculty share, there is still substantial gender-based sorting within fields. And even after accounting for student gender, sexist attitudes predict a significantly lower faculty share.

Fortunately, the methodology that we adopt is helpful for addressing the disadvantage that female faculty face. Our findings inform policy-relevant solutions which can range from the complex to the relatively simple. A complex solution would be to provide ratings for female and male faculty that adjust for gender-specific generosity and are normed to be equivalent across genders. This is technically feasible but sacrifices transparency. A simpler solution would flag to administrators courses in which female faculty face an expected disadvantage — either based on the gender composition of the course or the gender-specific generosity of the students. Our methodology informs the necessary adjustments depending on the specific context and provides policy makers with a tool to combat the systematic disadvantage experienced by female faculty.

References

Abowd, J. M., Kramarz, F., and Margolis, D. N. (1999). High wage workers and high wage firms. *Econometrica*, 67(2):251–333.

Acosta-Soto, L., Okoye, K., Camacho-Zuñiga, C., Escamilla, J., and Hosseini, S. (2022). An analysis of the students' evaluation of professors' competencies in the light of professors' gender. In *2022 IEEE Frontiers in Education Conference (FIE)*, pages 1–7.

Amer, A., Craig, A., and Effenterre, C. V. (2024). Decoding gender bias: The role of personal interactions. *Working paper*.

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. *SSRN Working paper*.

Ayllón, S. (2022). Online teaching and gender bias. *Economics of Education Review*, 89:102280.

Baker, R., Dee, T., Evans, B., and John, J. (2022). Bias in online classes: Evidence from a field experiment. *Economics of Education Review*, 88:102259.

Beaman, L., Chattopadhyay, R., Duflo, E., Pande, R., and Topalova, P. (2009). Powerful Women: Does Exposure Reduce Bias?*. *The Quarterly Journal of Economics*, 124(4):1497–1540.

Becker, G. S. (1971). *The Economics of Discrimination, 2nd Edition*. University of Chicago Press, Chicago, IL.

Binderkrantz, A., Bisgaard, M., and Lassesen, B. (2022). Contradicting findings of gender bias in teaching evaluations: evidence from two experiments in denmark. *Assessment & Evaluation in Higher Education*, 47(8):1345–1357.

Bohren, J. A., Imas, A., and Rosenberg, M. (2019). The dynamics of discrimination: Theory and evidence. *American Economic Review*, 109(10):3395–3436.

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

Boring, A., Ottoboni, K., and Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*.

Card, D., DellaVigna, S., Funk, P., and Iriberri, N. (2019). Are Referees and Editors in Economics Gender Neutral?*. *The Quarterly Journal of Economics*, 135(1):269–327.

Carrell, S. E., Page, M. E., and West, J. E. (2010). Sex and science: How professor gender perpetuates the gender gap. *The Quarterly journal of economics*, 125(3):1101–1144.

Eberhardt, M., Facchini, G., and Rueda, V. (2023). Gender Differences in Reference Letters: Evidence from the Economics Job Market. *The Economic Journal*, 133(655):2676–2708.

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

Glick, P. and Fiske, S. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, 70(3):491–512.

Goldin, C. and Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, 90(4):715–741.

Keng, S.-H. (2020). Gender bias and statistical discrimination against female instructors in student evaluations of teaching. *Labour Economics*, 66:101889.

MacNell, L., Driscoll, A., and Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40(4):291–303.

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

Mitchell, K. M. and Martin, J. (2018). Gender bias in student evaluations. *PS: Political Science & Politics*, 51(3):648–652.

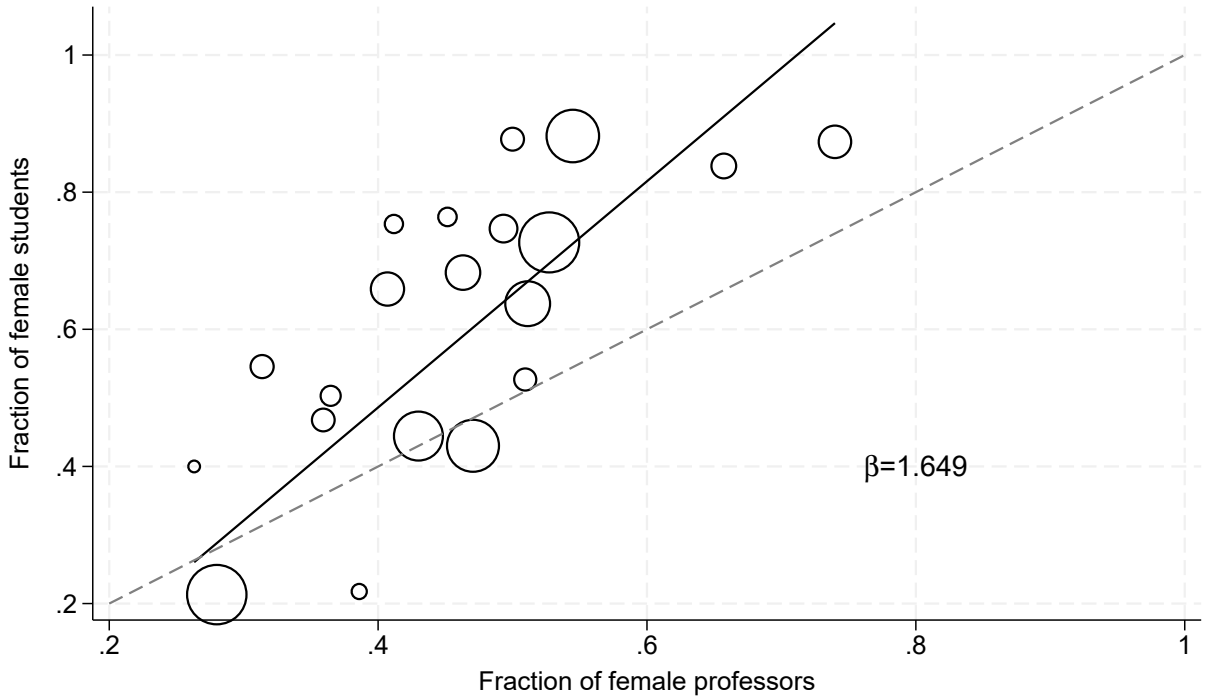
Morris, C. N. (1983). Parametric empirical bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):47–55.

Wagner, N., Rieger, M., and Voorvelt, K. (2016). Gender, ethnicity and teaching evaluations: Evidence from mixed teaching teams. *Economics of Education Review*, 54:79–94.

Wigert, B. and Harter, J. (2017). Re-engineering performance management. *Gallup. com. Viewed: March*, 6:2019.

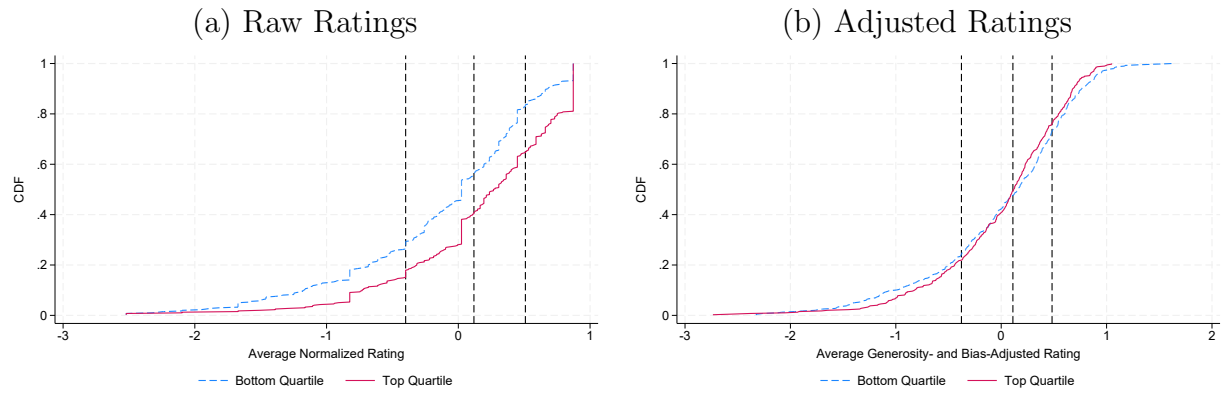
Figures

Figure 1: Fraction of Female Students and Female Professors



Notes: This figure plots the relationship between the fraction of female faculty within a degree program and the fraction of female students within a college major group. College major groups include: Architecture, Arts and Communications, Biology, Chemistry, Criminology, Economics, Education, Engineering, Geography, History, Law, Medicine, Nursing, Languages, Philosophy, Physical Therapy, Political Science, Psychology, Social Work, and Marketing.

Figure 2: Calculating Student Disadvantage



Notes: Panel A plots cumulative density functions of average normalized ratings for female faculty with bottom- and top-quartile draws of predicted gender-specific generosity. Panel B plots cumulative density functions of ratings that have been adjusted for instructor draws of gender-specific generosity and for major-specific gender bias for female faculty with bottom- and top-quartile draws of predicted gender-specific generosity.

Tables

Table 1: Predicting Female Faculty Share, Survey Evidence

	(1)	(2)	(3)	(4)	(5)	(6)
Female Student	0.135***	0.105***			0.121***	0.091***
	(0.027)	(0.027)			(0.027)	(0.027)
Sexism Measure			-0.043***	-0.042***	-0.029**	-0.032**
			(0.014)	(0.014)	(0.014)	(0.014)
Observations	359	359	359	359	359	359
R^2	0.066	0.207	0.027	0.194	0.078	0.220
Major FE	–	X	–	X	–	X

Notes: Observations are at the student level. The outcome is the fraction of a student's five most recent courses that were taught by a female instructor. Our sexism measure is constructed from four externally validated gender attitude questions that ask students how much they agree with the following statements: (1) No matter how accomplished he is, a man is not truly complete until he has the love of a woman. (2) Many women are actually seeking special favors, such as hiring policies that favor them over men, under the guise of asking for equality. (3) Women are too easily offended. (4) Many women have a quality of purity that few men possess. Significance levels: $*$: 10% $**$: 5% $***$: 1%.

Table 2: Evidence of Potential Gender Bias

	Student Rating				
	(1)	(2)	(3)	(4)	(5)
Instructor Female	0.015	-0.011	-0.013	-0.042*	-0.046**
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
Obs	263,460	263,460	263,460	263,460	263,460
R^2	0.000	0.008	0.012	0.028	0.051
Faculty Characteristics		X	X	X	X
Field and Course Characteristics			X	X	X
Student Characteristics				X	X
Student Final Grade					X

Notes: Coefficients show regression results of normalized student ratings on an indicator variable for whether the professor is female. These regressions reflect professor-level results as observations are weighted by the inverse of the number of student responses for the professor. Faculty controls include professor age and rank. "Faculty characteristics" include lecturer's age fixed effects and tenure ("Full professor", "Associate professor", "Assistant professor" or "Visiting professor" and "Other" — typically pre-doctoral students and adjunct faculty); "Field and course characteristics" include field of study, elective or mandatory course, fixed-effects by academic semester; "Student characteristics" include student gender, student age fixed effects, course repeater and degree; finally, "Student final grade" refers to the overall grade obtained at the end of the semester for a given course. Standard errors clustered at the professor level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 3: Distribution and Predictors of Student Generosity and Bias

Panel A: Standard Deviations of Student Generosity and Bias

	Overall Generosity	Generosity to Male Instructors	Generosity to Female Instructors	Bias
	(1)	(2)	(3)	(4)
SD of Empirical Measure	0.462	0.605	0.529	0.609
SD of Latent Measure	0.340	0.351	0.367	0.207

Panel B: Predictors of Student Generosity and Bias

	(1)	(2)	(3)	(4)
Student Female	0.046***	0.012	0.087***	-0.075***
	(0.010)	(0.011)	(0.013)	(0.013)
Student Age	0.012***	0.011***	0.015***	-0.004***
	(0.001)	(0.001)	(0.001)	(0.001)
Obs	12,468	12,468	12,468	12,468

Notes: In Panel A The first row of results shows the standard deviation of empirical measures of generosity from a two-way fixed effects regression of student ratings on faculty fixed effects, student fixed effects, course year, and an indicator for spring semester performed separately by student major. The sample includes students who rated at least one male and one female faculty member. Bias is measured as the difference between male and female generosity. The second row shows the standard deviation of the latent measures of generosity and bias calculated as described in the text. For Panel B * p<0.10, ** p<0.05, *** p<0.01. The dependent variable of these regressions are the student-level generosity and bias measures estimated from a two-way fixed effects regression of student ratings on faculty fixed effects, student fixed effects, and course semester performed separately by student major.

Table 4: Forecasting Generosity and Bias

	Predicting Fixed Effect				Predicting Individual Rating		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Overall Generosity	0.988***				1.006***		
	(0.041)				(0.038)		
Generosity towards Males		1.007***				0.966***	
		(0.051)				(0.045)	
Generosity towards Females			0.905***				0.942***
			(0.057)				(0.045)
Bias				0.493***			
				(0.130)			
P-value (=1)	.779	.894	.092	0	.867	.445	.195
Obs	4,378	3,747	3,044	2,595	43,302	24,924	18,378

Notes: This regression tests whether shrunken measures of generosity and bias predict future generosity and bias out of sample. The shrunken measures are calculated using data prior to 2021 as described in Appendix section B. In columns 1-4, the dependent variable of these regressions are the student-level generosity and bias measures estimated from a two-way fixed effects regression of student ratings on faculty fixed effects, student fixed effects, and an indicator for spring semester performed separately by student major using 2021 data. In columns 5-7, the dependent variable is a normalized student rating. Controls include course fixed effects, an indicator for spring semester and major fixed effects. All hypothesis tests are conducted relative to a null hypothesis that the coefficient on the shrunken measure is 1. In column 1, robust standard errors are shown. In column 2, standard errors are cluster-corrected at the student level. * p<0.10, ** p<0.05, *** p<0.01.

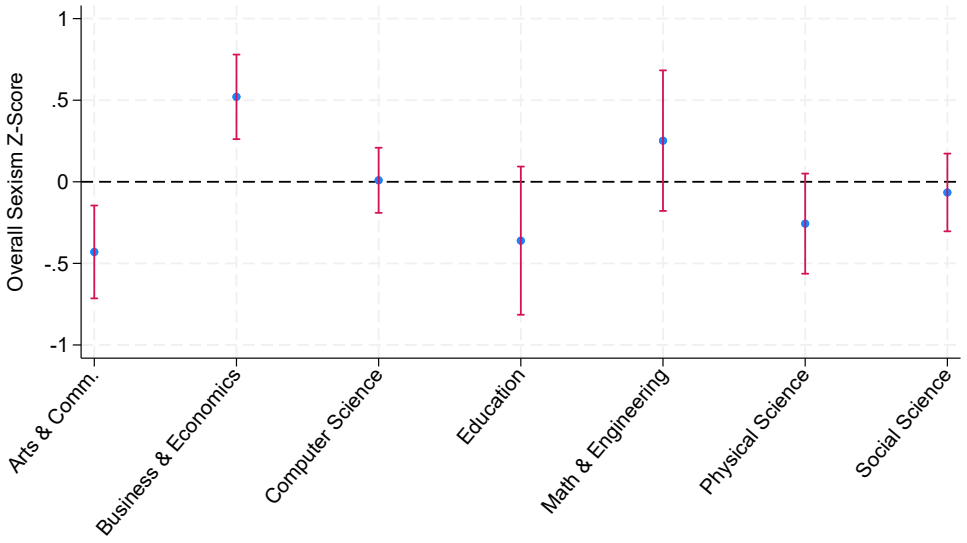
Table 5: Identifying Bias and Predicting Student Sorting

	Student Rating		Faculty FE		Student FE (Generosity)		Bias		Fraction Female Students	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Professor Female	-0.010	-0.036*	-0.018	-0.044**	0.013**	0.013**	-0.009	-0.013	0.035***	0.034***
	(0.019)	(0.019)	(0.019)	(0.019)	(0.006)	(0.006)	(0.011)	(0.011)	(0.005)	(0.005)
Professor Age		-0.010***		-0.009***		-0.000		-0.002***		0.000
		(0.001)		(0.001)		(0.000)		(0.001)		(0.000)
Professor Permanent		0.010		-0.011		-0.005		0.053***		-0.007
		(0.024)		(0.024)		(0.007)		(0.013)		(0.006)
Obs	3,099	3,099	3,098	3,098	3,098	3,098	3,084	3,084	3,099	3,099
Degree Fixed Effects	X	X	X	X	X	X	X	X	X	X

Notes: Robust standard errors in parentheses. Regressions are at the professor-degree level. The professor-level fixed effects, the student-level generosity, and bias measures are estimated from a two-way fixed effects regression of student ratings on faculty fixed effects, student fixed effects, and course semester performed separately by student major. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

A Appendix

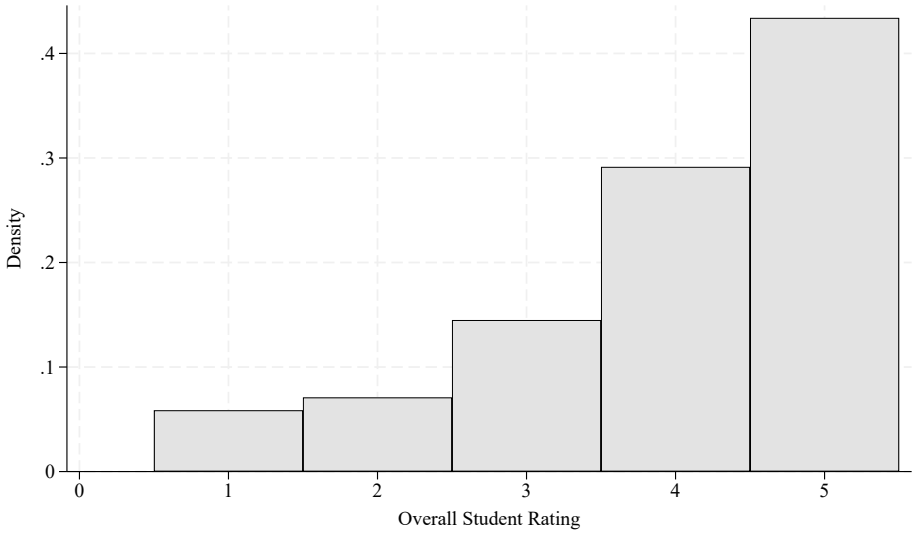
Figure A.1: Average Sexism by Major Field, Survey Evidence



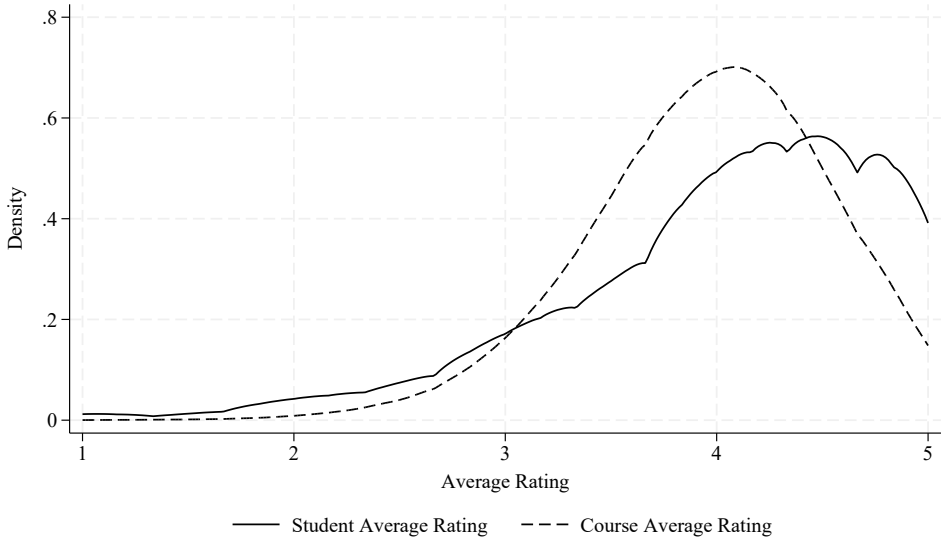
Notes: This figure reports variation in sexist attitudes by degree field from a survey of U.S. undergraduates conducted by the authors. Our sexism measure is constructed from summing and normalizing responses to four externally validated gender attitude questions. These questions ask students how much they agree with the following statements: (1) No matter how accomplished he is, a man is not truly complete until he has the love of a woman; (2) Many women are actually seeking special favors, such as hiring policies that favor them over men, under the guise of asking for equality; (3) Women are too easily offended; and, (4) Many women have a quality of purity that few men possess.

Figure A.2: Student Ratings of Instructor Overall Performance

(a) Individual Ratings

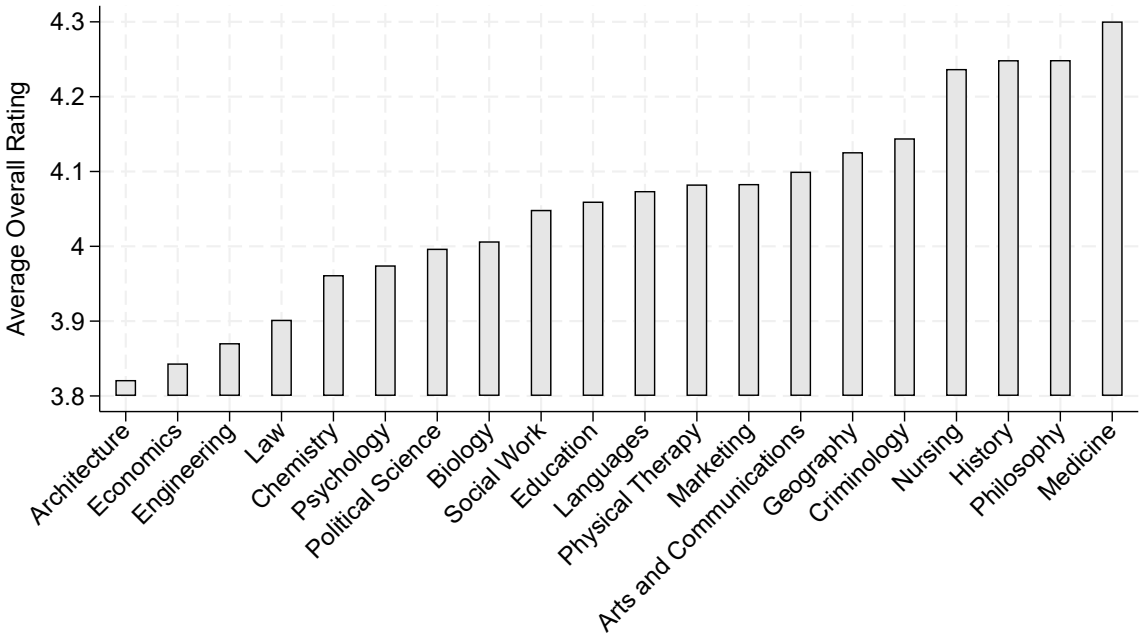


(b) Average Ratings



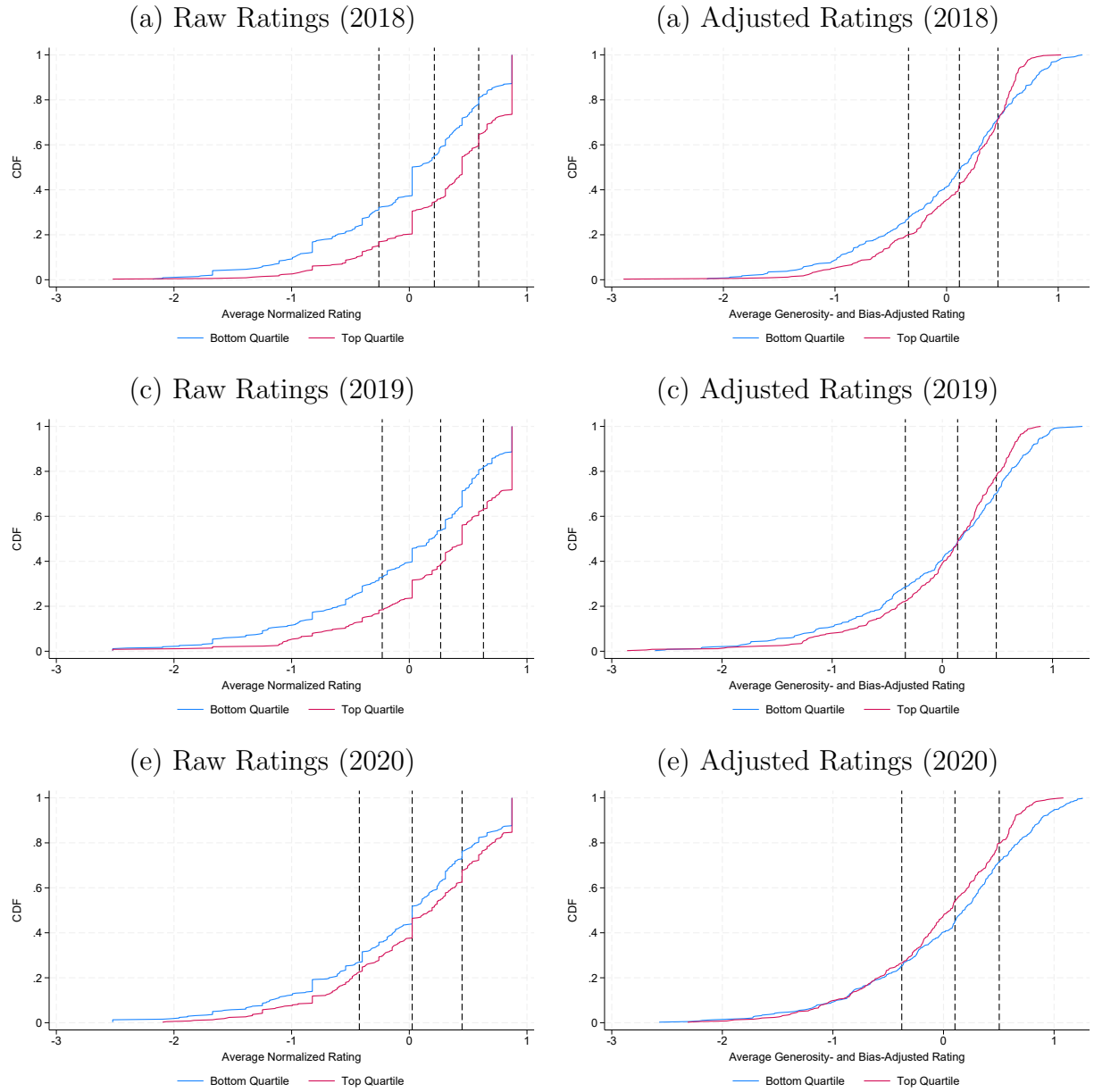
Notes: Panel A shows the distribution of individual student ratings at the University of Girona. Panel B plots the CDFs of student average ratings and course average ratings.

Figure A.3: Student Ratings of Instructor Overall Performance, By Major



Notes: Each bar represents the average rating in courses taught within the corresponding major.

Figure A.4: Calculating Student Disadvantage In Different Years



Notes: Panels A, C, and E plot cumulative density functions of average normalized ratings for female faculty with bottom- and top-quartile draws of predicted gender-specific generosity. Panels B, D, and F plot cumulative density functions of ratings that have been adjusted for instructor draws of gender-specific generosity and for major-specific gender bias for female faculty with bottom- and top-quartile draws of predicted gender-specific generosity.

Table A.1: Summary Statistics

	All	Est. Sample	Male	Female	P-value (3) vs. (4)
Student Female	0.593	0.595	0.000	1.000	-
	(0.491)	(0.491)	(0.000)	(0.000)	-
Student Age	21.684	21.564	21.695	21.474	-
	(4.356)	(4.156)	(4.256)	(4.085)	0.001
Student GPA	6.889	6.888	6.510	7.145	-
	(1.340)	(1.312)	(1.350)	(1.221)	0.000
Student Repeats Course	0.007	0.007	0.009	0.005	-
	(0.082)	(0.082)	(0.096)	(0.070)	0.001
Professor Female	0.421	0.420	0.338	0.476	-
	(0.493)	(0.493)	(0.473)	(0.499)	0.000
Professor Age	47.500	47.471	47.814	47.237	-
	(9.647)	(9.611)	(9.633)	(9.589)	0.000
Professor Permanent	0.464	0.465	0.521	0.427	-
	(0.497)	(0.497)	(0.498)	(0.492)	0.000
Arts and Humanities	0.074	0.063	0.060	0.064	-
	(0.261)	(0.242)	(0.238)	(0.245)	0.331
Sciences	0.111	0.116	0.109	0.121	-
	(0.314)	(0.321)	(0.311)	(0.327)	0.014
Health Sciences	0.163	0.163	0.144	0.175	-
	(0.369)	(0.369)	(0.351)	(0.380)	0.000
Social Sciences	0.471	0.476	0.346	0.564	-
	(0.499)	(0.499)	(0.476)	(0.496)	0.000
Engineering and Architecture	0.182	0.183	0.340	0.076	-
	(0.386)	(0.387)	(0.474)	(0.264)	0.000
Mandatory Course	0.880	0.894	0.914	0.880	-
	(0.324)	(0.308)	(0.280)	(0.325)	0.000
Instructor Motivates	3.945	3.937	3.871	3.982	-
	(0.744)	(0.744)	(0.781)	(0.714)	0.000
Instructor is Helpful	3.766	3.756	3.696	3.797	-
	(0.796)	(0.794)	(0.835)	(0.762)	0.000
Overall Student Rating	4.124	4.116	4.061	4.153	-
	(0.775)	(0.779)	(0.823)	(0.745)	0.000
Observations	17780	15862	6421	9441	-

Notes: Each observation corresponds to a single student. Variables computed as the average for each student across all the surveys a student answers. Standard errors in parenthesis. Field of study comes from the student rather than course. Column 1 includes all observations. Columns 2-4 use the sample for which all the analysis of the paper are conducted: students who have a stated major and are not enrolled in a small specialty program.

Table A.2: Forecasting Generosity and Bias In Different Years

	Predicting Fixed Effect				Predicting Individual Rating		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
2018							
Overall Generosity	1.087***				1.125***		
	(0.049)				(0.042)		
Generosity towards Males		0.973***				1.014***	
		(0.056)				(0.052)	
Generosity towards Females			1.001***				1.004***
			(0.066)				(0.054)
Bias				-0.017			
				(0.195)			
P-value (=1)	.077	.627	.993	0	.003	.78	.947
Obs	4,559	3,702	3,126	2,515	34,540	19,464	15,076
2019							
Overall Generosity	0.793***				0.843***		
	(0.045)				(0.041)		
Generosity towards Males		0.720***				0.790***	
		(0.048)				(0.040)	
Generosity towards Females			0.928***				0.805***
			(0.065)				(0.059)
Bias				1.156***			
				(0.129)			
P-value (=1)	0	0	.267	.227	0	0	.001
Obs	4,340	3,532	2,930	2,374	32,841	18,122	14,719
2020							
Overall Generosity	1.071***				1.095***		
	(0.050)				(0.047)		
Generosity towards Males		0.917***				1.042***	
		(0.060)				(0.054)	
Generosity towards Females			1.038***				0.960***
			(0.077)				(0.059)
Bias				0.313*			
				(0.138)			
P-value (=1)	.156	.166	.621	0	.042	.441	.498
Obs	4,361	3,752	2,989	2,569	42,670	25,115	17,555
2021							
Overall Generosity	0.989***				1.006***		
	(0.041)				(0.038)		
Generosity towards Males		1.007***				0.966***	
		(0.051)				(0.045)	
Generosity towards Females			0.905***				0.942***
			(0.057)				(0.045)
Bias				0.493***			
				(0.130)			
P-value (=1)	.78	.894	.092	0	.868	.445	.195
Obs	4,378	3,747	3,044	2,595	43,302	24,924	18,378

Standard errors clustered at the student level. * p<0.10, ** p<0.05, *** p<0.01. The dependent variable is whether an instructor is female. The sample includes students from 2021. The overall generosity and bias measures are shrunk and constructed as described on page XX.

B Modified Empirical Bayes Procedure for Estimating Predictions of Generosity and Bias

The degree of predictability of our measures of generosity and bias is substantially reduced by estimation error. A typical way to overcome this challenge is to construct an empirical Bayes (EB) measure that shrinks the noisy measure closer to the conditional mean as described by Morris (1983). The implementation of the EB method is complicated by the fact that it is a challenge to construct valid standard errors for tens of thousands of fixed effects. Additionally, students who are generous towards male faculty tend to be generous towards female faculty as well. This correlation causes problems for calculating standard EB measures of gender-specific generosity and bias. Given these challenges, we implement the following method for constructing predictions of generosity and bias that are approximately forecast unbiased.

For simplicity, we first describe our method for predicting overall generosity. We first estimate Equation 4 using the observations from the year 2015 to 2019. We call this sample A. We then estimate Equation 4 using only observations from the year 2020, which we refer to as sample B. We then run the following student-level regression:

$$\hat{\phi}_i^B = \gamma_0 + \gamma_1 \hat{\phi}_i^A + \gamma_2 \frac{\hat{\phi}_i^A}{N_i^A} + \gamma_3 \frac{1}{N_i^A} + X_i \Gamma + u_i \tag{1}$$

This equation shows how student level covariates, X_i, and our raw measure of generosity from sample A, $\hat{\phi}_i^A$, predicts out-of-sample generosity, $\hat{\phi}_i^B$. We interact $\hat{\phi}_i^A$ with the inverse of the number of evaluations completed by student i in sample A, which takes into account that the variance of $\hat{\phi}_i^A$ is roughly proportional to $\frac{1}{N_i^A}$. The coefficient, γ_2, allows the predictive power of the raw measures to increase with the precision of the estimate in a manner similar to the standard EB method. The estimated parameters of this model allow us construct a "best guess" of actual generosity that will be close to forecast unbiased if

the data generating process of student evaluations is stationary.

To test the performance of these estimates out of sample, we calculate raw measures of generosity by estimating equation 4 using data from 2015 to 2020, which we call sample C. We then use the parameter estimates from equation 1 to create our best guesses of actual student generosity, $\tilde{\phi}_i^C$, in the following manner.

$$\tilde{\phi}_i^C = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{\phi}_i^C + \hat{\gamma}_2 \frac{\hat{\phi}_i^C}{N_i^C} + \hat{\gamma}_3 \frac{1}{N_i^C} + X_i \hat{\Gamma} + u_i \tag{2}$$

We use student ratings from the year, 2021, as an evaluation data set, which we denote as data set D. We estimate 4 with just these observations to construct measures of student-level generosity, $\hat{\phi}_i^D$. We then predict $\hat{\phi}_i^D$ using $\tilde{\phi}_i^C$ by estimating the equation:

$$\hat{\phi}_i^D = \alpha_0 + \alpha_1 \tilde{\phi}_i^C + e_i \tag{3}$$

If our "best guess" of student generosity is predictive out-of-sample, we would expect α_1 to be positive and significant. If it is forecast unbiased, we would expect α_1 to be insignificantly different from 1. We can also test the out-of-sample performance of these measures using student microdata by regressing individual student level rating from sample D on $\tilde{\phi}_i^C$ along with professor and degree group fixed effects.

The process for estimating gender-specific generosity measures is the same except for the fact that one needs to take into account that generosity towards faculty of one gender is predictive of generosity towards faculty of the other gender. Consequently, the analog to Equation 1 for estimating a "best guess" of generosity towards female faculty is given by:

$$\hat{\phi}_{fi}^B = \gamma_f 0 + \gamma_{f1} \hat{\phi}_{fi}^A + \gamma_{f2} \frac{\hat{\phi}_{fi}^A}{N_{fi}^A} + \gamma_{f3} \frac{1}{N_{fi}^A} + \gamma_{f4} \hat{\phi}_{mi}^A + \gamma_{f5} \frac{\hat{\phi}_{mi}^A}{N_{mi}^A} + \gamma_{f6} \frac{1}{N_{mi}^A} + X_i \Gamma_f + u_i \tag{4}$$

The subscripts in this equation denote the gender of the professor. The model allows generosity towards male and female professors to have independent predictive ability for

future generosity towards female faculty. The coefficients have gender subscripts because the coefficients are likely to differ when predicting generosity towards male faculty. Once we estimate Equation 4, we construct $\tilde{\phi}_{if}^{C}$ in a manner analogous to what we did for overall generosity. We can construct similar measures for generosity towards male faculty. Our "best guess" for bias is given simply by $\tilde{bias}_i^C = \tilde{\phi}_{im}^{C} - \tilde{\phi}_{if}^{C}$.

C Survey

During the summer of 2023 we administered a survey on Prolific to 359 college students in the U.S. who were enrolled at a four-year college or university and had taken at least four classes during the previous six months. The survey was approved by the BYU IRB (IRB2023-158) and took about 15 minutes to complete on average.

After collecting informed consent and demographic information, we asked respondents to identify four specific classes they had taken most recently. We then asked them to rank the classes from worst to best based on the following criteria: overall ranking, alignment with student's interests, usefulness to the student's chosen career path or field of subsequent study, difficulty, and the time of instruction. Respondents also indicated whether each class was required for their major or general education and the grade that they received in each course.

In the next section of the survey, we asked students to rank instructors for these classes from worst to best based on the following criteria: overall effectiveness, ability to explain difficult concepts, organizational skills, kindness and caring personality, competence, and time commitment from the students. We also asked respondents to indicate whether each instructor was a permanent or adjunct faculty as well as their perceived age, gender, and race.

Finally, we asked respondents to state the degree to which they agree or disagree with the four statements from the standard ambivalent sexism scale (Glick and Fiske, 1996) to measure students' gender attitudes. Two of the statements were used to measure benevolent sexism and two for hostile sexism. We outline the survey protocol in sections C1-C7 below.

Our respondents are on average 29 years old. The majority (63%) are currently enrolled in

a degree program at a four-year college or university in the U.S. while 37% have graduated within the last six months. About half (52%) are college juniors or seniors, 45% are women, 50% are White, 20% are Black, 12% are Asian, and 15% are Hispanic. 63% of respondents self-identify as strongly or moderately liberal on most political matters. An average respondent took 15.6 minutes to complete the survey as was paid $3.5 for their participation. 99.7% of subjects passed the attention check.

C.1 Screening Questions

Are you currently in the United States?

Are you at least 18 years old?

Are you currently a student at a four-year college or university?

Have you taken at least four different classes at a four-year college or university over the last six months?

C.2 Consent to Participate in a Research Study

Title of the Study: Student Survey

Principal Investigator: Olga Stoddard (Brigham Young University)

Phone: 801-574-3014

Email: olga.stoddard@byu.edu

You are being asked to volunteer in a research study. Below, you will find information about this research for you to carefully consider when deciding about whether or not to participate. Please ask questions about any of the information you do not understand before you decide whether to participate.

Key Information for You to Consider

Statement of Research: Purpose. The purpose of this research is to learn more about

decision-making in college. You are being asked to volunteer for a research study. It is up to you whether you choose to participate or not. There will be no penalty or loss of benefits to which you are otherwise entitled if you choose not to participate or discontinue participation.

Duration. It is expected that your participation will last 10 minutes.

Procedures and Activities. You will be asked to fill out a survey.

Risks: We do not believe there are any reasonably foreseeable risks, discomforts, hazards or inconveniences for participants for participation in this research.

Benefits: There may be no personal benefit from your participation but the knowledge received may be of value to humanity.

What is this study about? Researchers at Brigham Young University are conducting a study on students' academic experiences in a variety of university and college classes. You are being asked to participate because we believe you are currently taking university/college classes as a student. Your participation in the study is expected to last 10 minutes. The study is supported by Brigham Young University.

What will happen during this research? If you agree to participate in this research, your participation will include filling out a survey and having your responses reported on when aggregated with other's responses in research materials.

The information collected as part of this research will not be used or distributed for future research studies, even if all of your identifiers are removed. We will tell you about any new information that may affect your willingness to continue participation in this research.

What are the risks or discomforts associated with this research? We do not

believe there are any reasonably foreseeable risks, discomforts, hazards or inconveniences for participants for participation in this research.

How might I benefit from this research? There may be no personal benefit from your participation.

What is the compensation for the research? If you complete the entire survey, you will receive the compensation advertised to you on the platform where you found this opportunity.

What will happen if I choose not to participate? It is your choice to participate or not to participate in this research. Participation is voluntary. Alternatives to participation are leaving this webpage.

Is my participation voluntary, and can I withdraw? Taking part in this research study is your decision. Your participation in this study is voluntary. You do not have to take part in this study, but if you do, you can stop at any time by leaving this webpage. Your decision whether to participate will not affect your relationship with the researchers or their organizations. There are no penalties/consequences/loss of benefits to which you are otherwise entitled, if you do not participate. However, you will not be paid the compensation advertised to you on the platform where you found this opportunity if you do not complete the survey.

You have the right to choose not to participate in any study activity or completely withdraw from continued participation at any point in this study without penalty/consequences/loss of benefits to which you are otherwise entitled. If you withdraw from the study, the data collected to the point of withdrawal will be deleted.

Who do I talk to if I have questions?

If you have questions, concerns, or have experienced a research-related injury, contact the research team at:

Dr. Olga Stoddard

801-422-3580

olga.stoddard@byu.edu

An Institutional Review Board ("IRB") is overseeing this research. IRB is a group of people who perform independent review of research studies to ensure the rights and welfare of participants are protected. If you have questions about your rights or wish to speak with someone other than the research team, you may contact:

Brigham Young University IRB

(801) 422-3606

irb@byu.edu

Statement of Consent I have read and considered the information presented in this form. I confirm that I understand the purpose of the research and the study procedures. I understand that I may ask questions at any time and can withdraw my participation without prejudice. I have read this consent form. By clicking on the arrow button to continue I indicate my willingness to participate in this study.

C.3 Demographics

Please answer the following questions about yourself.

1. What is your age?
2. When do you expect to complete your degree? (Enter year. - eg., 2025)
3. What is your GPA? (Enter a number between 0 and 4.0)
4. What is your gender?
5. What ethnic group do you belong to?
6. On most political matters do you consider yourself: o Strongly conservative o

Moderately conservative o Neither, middle of the road o Moderately liberal o Strongly liberal o Prefer not to state

7. The following question about your hobbies is very simple. When asked what you are doing, please select "I am running" from the options below, no matter what you are actually doing right now. This is an attention check. Based on the instructions above, what hobby have you been asked to select? o I am swimming o I am running o I am taking a survey o I am playing the piano

8. What is the highest level of education you have achieved? o Some high school o High school diploma or equivalent o Some college o Associate's degree o Bachelor's degree o Graduate Degree

9. What is your major (if you have multiple majors, list them all; if undecided, state "undecided")

10. Over the last year, during which of the following semesters were you enrolled in classes? (Check all that apply)

11. How many classes did you take in the (insert the first semester that they selected above in chronological order (i.e. least recent)?

12. How many classes did you take in the (insert the second semester that they selected above)?

13. How many classes are you taking currently? (only include if they are "currently enrolled")

C.4 Classes

Next, please identify the four courses you took the Spring 2023 semester:

1. Think of the first class you attended each week. For example, this might be a Monday morning class. What is the catalog number of this class (e.g. CHEM 100)? If you don't remember, just give it your best guess. This allows us to have a convenient indicator of the class for future questions.

2. Think of the second class you attended each week. What is the catalog number of this class (e.g. CHEM 100)? If you don't remember, just give it your best guess. This allows us to have a convenient indicator of the class for future questions.
3. Think of the third class you attended each week. What is the catalog number of this class (e.g. CHEM 100)? If you don't remember, just give it your best guess. This allows us to have a convenient indicator of the class for future questions.
4. Think of the fourth class you attended each week. What is the catalog number of this class (e.g. CHEM 100)? If you don't remember, just give it your best guess. This allows us to have a convenient indicator of the class for future questions.

C.5 Class Rankings

Please rank these four classes from worst to best ('1' corresponding to worst and '4' to best) based on the following dimensions:
1. Which class was best overall?
2. Which class was most closely linked to your interests?
3. Which class was most useful either for your career or for subsequent study?
4. Which class was most difficult?
5. Which class was taught at the best time?
6. Were any of these classes part of a general education requirement?
7. Were any of these classes required for your major?
8. What grade did you receive in CLASS 1? (enter letter grade A, B, C, D, F, or NA, allowing for + and - (eg. B+))
9. What grade did you receive in CLASS 2? (enter letter grade A, B, C, D, F, or NA, allowing for + and - (eg. B+))
10. What grade did you receive in CLASS 3? (enter letter grade A, B, C, D, F, or NA, allowing for + and - (eg. B+))
11. What grade did you receive in CLASS 4? (enter letter grade A, B, C, D, F, or NA,

allowing for + and - (eg. B+))

C.6 Instructor Rankings

You will now be asked to rank the instructors for these four classes from worst to best (1 corresponding to worst and 4 to best) based on the following dimensions:

1. Which instructor was overall most effective?
2. Which instructor did you like best?
3. Which instructor was best at explaining challenging concepts?
4. Which instructor was most interesting or engaging?
5. Which instructor was most organized?
6. Which instructor was most caring and kind?
7. Which instructor seemed to have the best command of the course material?
8. Which instructor demanded the most in terms of time commitment from the students?
9. Was the instructor of CLASS 1 an adjunct or permanent professor?
10. Was the instructor of CLASS 2 an adjunct or permanent professor?
11. Was the instructor of CLASS 3 an adjunct or permanent professor?
12. Was the instructor of CLASS 4 an adjunct or permanent professor?
13. Approximately how old was the instructor of CLASS 1?
14. Approximately how old was the instructor of CLASS 2?
15. Approximately how old was the instructor of CLASS 3?
16. Approximately how old was the instructor of CLASS 4?
17. What was the gender of the instructor of CLASS 1?
18. What was the gender of the instructor of CLASS 2?
19. What was the gender of the instructor of CLASS 3?
20. What was the gender of the instructor of CLASS 4?
21. What was the race of the instructor of CLASS 1?
22. What was the race of the instructor of CLASS 2?

23. What was the race of the instructor of CLASS 3?

24. What was the race of the instructor of CLASS 4?

C.7 Gender Attitudes

Below is a series of statements concerning men and women and their relationship in contemporary society. Please indicate the degree to which you agree or disagree with each statement.

1. No matter how accomplished he is, a man is not truly complete until he has the love of a woman.

2. Many women are actually seeking special favors, such as hiring policies that favor them over men, under the guise of asking for equality.

3. Women are too easily offended.

4. Many women have a quality of purity that few men possess.