Gender Bias in Performance Evaluations: Evidence from a Field Experiment[*]

Perihan O. Saygin[†]
Thomas Knight[‡]

January 20, 2023

Abstract

Despite a shrinking or even reversed gap in educational attainment, women continue to be underrepresented in leadership positions, and a gender wage gap persists. Performance evaluations are a primary determinant of hiring and promotion decisions, and peer evaluations have become increasingly widespread in the workplace. There is, however, little evidence of whether peer evaluators exhibit gender bias. We identify and measure bias in performance evaluations in a large, introductory course at a flagship public university. Peer evaluators were randomly assigned to score essays using a rubric. Evaluators were incentivized to match official grades, adding a monitoring effect. We exploit the random assignments of both peer evaluators and blinded official graders over several essay assignments. Using student and peer evaluator fixed effects and conditioning on blinded official grades, we find that male peer graders assign higher scores to classmates without female-sounding names, relative to those with female sounding names, and that students without a female-sounding name receive higher scores when they are randomly assigned to a male peer grader as opposed to a female grader. We do not find such biases for women. The observed biases could be even more pronounced in real-world settings, where monitoring is less common. These results suggest that biased performance evaluations could be at least partly responsible for gender gaps in hiring and promotion, particularly in male-dominated fields where evaluators are more likely to be men.

JEL Classification: J16, I23
Keywords: performance evaluation, peer evaluation, gender bias.

[*] We would like to thank Robert Ainsworth, Drew Atchison, David Card, Damon Clark, Chris Curran, Maria A. Davia, Mark Hoekstra, Nagore Iriberri, Scott Kostyshak, Peter Orazem, Maria Padilla-Romo, Patricia Palffy, Garrison Pollard, Andrew Rollins, Rich Romano, David Sappington, Mark Rush, Sarah Turner, Zhenda Yin for thoughtful comments. We also would like to thank AEDE 2021, MEA-SOLE 2021, SMLW 2021, SOLE 2021, WATE 2021 conferences, and Armenian Economic Association and TOBB Economics seminars participants. All errors are our own. IRB ID: 202100434 (Study: Discrepancies between Peer Assigned and Instructor Assigned Grades). Corresponding author's email: psaygin@ufl.edu

[†] Department of Economics, Robert F. Lanzillotti Public Policy Research Center, affiliate faculty, Education Policy Research Center and Center for Gender, Sexualities, and Women's Studies Research, University of Florida.

[‡] Department of Economics, Robert F. Lanzillotti Public Policy Research Center, University of Florida.

1 Introduction

Women still struggle to reach the top of the corporate, academic, and political ladders despite the reversed gender gap in educational attainment that has occurred over the past three decades. While the share of women CEOs at Fortune 500 companies is at an all-time high, it currently sits at an underwhelming 9%. Moreover, only one in four C-Suite executives is a woman. For every 100 men who are promoted from an entry-level to managerial position, only 87 women experience similar promotion [LeanIn.Org and Company, 2022].

One potential driver of the observed gender gap in leadership positions is bias in performance evaluations. Evaluations often serve a critical role in determining promotions and raises for employees, as well as compensation and options packages for executives. Hiring and promotion decisions are often and increasingly made based on peer reviews (e.g., interviews). Several large employers, including Amazon, Google, Netflix, and Spotify, have recently moved away from traditional annual performance evaluations and replaced them with more frequent peer evaluations[1]. Additionally, many employers rely on peer referrals to reduce the cost of recruiting and screening applicants. Despite the rising importance of peer evaluations on career outcomes, there is a lack of evidence regarding the existence and magnitude of gender biases in these evaluations. Moreover, isolating the gender bias in evaluations for equally performing individuals is challenging.

In this paper, we aim to fill this gap in an experimental setting by analyzing the results of a peer grading scheme in a large, mostly online introductory course at a flagship public research university. We compare peer reviews of four short essay assignments to teaching assistant (TA) evaluations of the same submissions. Our dataset consists of 2891 homework submissions that were evaluated by randomly assigned, blind TAs and randomly assigned, non-blind peer graders. The assignment grades are determined by TAs' evaluations. These scores can be decomposed into content and writing subscores. The degree of objectivity varies across these two subscores. The content subscore measures correctness of responses to specific questions in the assignment prompts with objectively correct answers. The writing subscore measures general writing quality. The peer grading setting that we examine offers a unique opportunity to investigate many relevant questions

[1] See the article Di Fiore and Marcio [2021] for a discussion of whether peer reviews are the future of performance evaluations.

about gender biases in evaluations: Does the gender of the peer evaluator matter? Does the peer evaluator's gender matter differentially when the assignment author has a female or male-sounding name? Do students with female-sounding names receive lower peer evaluations overall? Are these differences similar in content subscores and writing subscores? How does the academic performance of the peer evaluator affect how they evaluate their classmates' work?

We find that, overall, peer evaluators assign lower scores than TAs, and that this difference is more pronounced in the (more objective) content subscores than the (more subjective) writing subscores. We observe significant gender differences in peer evaluations, and these gender differences arise primarily through the content subscores.[2] Male peer evaluators assign 2.02-point higher scores, on average, to authors without a female-sounding name than they do to authors with a female-sounding name. Similarly, male authors receive 2.50-point higher scores, on average, when they are randomly assigned to a male peer evaluator as opposed to when they are assigned to a female evaluator. We do not find similar results for female peer evaluators or female authors. Female peer evaluators assign similar scores to authors with and without female-sounding names, and female students receive similar grades when they are randomly assigned to a male or female peer evaluator. Interestingly, these gender biases are observed in the content subscores, but the writing subscores do not exhibit similar biases.

Our paper examines an important source of gender bias in labor market outcomes. Specifically, it contributes to the literature on gender biases in performance evaluations which can explain the observed persistent gender disparities and underrepresentation of women in leadership positions and certain fields. Prior literature provides evidence for gender disparities in performance evaluations and referrals. Using content analysis of individual annual performance reviews, Cecchi-Dimeglio [2017] shows that women are 1.4 times more likely to receive critical subjective feedback, as opposed to either positive feedback or critical objective feedback. The data also revealed that women receive less constructive critical feedback. Correll and Simard [2016] suggest that women receive more vague feedback than men do. Additionally, Beaman et al. [2018], Sarsons [2017] and Zeltzer [2020] find consistent evidence that women receive fewer employment referrals.

[2] The relative influence of content subscores here is not driven by differential weighting in the assignment grade calculations, although this weighting does reinforce it. The observed gender differences in content subscores are independently much more pronounced than any differences in writing subscores.

Some studies provide evidence of bias by showing that women's qualifications and job performance are discounted, relative to that of their male counterparts, leading to gender disparities in performance evaluations and hiring outcomes [Goldin and Rouse, 2000, Milkman et al., 2015, Quadlin, 2018, Sarsons, 2017]. Our paper complements these studies. Most importantly, our paper quantifies gender biases in peer evaluations for equally performing individuals. We accomplish this by examining a unique setting in which the evaluators are randomly assigned, and an objectively measured performance indicator is available for comparison. Identifying and accurately measuring bias in performance evaluations is especially challenging, because evaluators are rarely randomly assigned. Even when evaluators are randomly assigned, there is rarely an objective performance measure against which to compare the evaluation[3]. In our setting, we overcome this obstacle with double randomization. We randomly assign students to evaluate their peers' work. We also randomly assign TA graders to blindly evaluate the same submissions. This experimental setting provides a clean objective performance measure for comparison: the randomly assigned TAs' blind scores of the same submissions[4].

Our findings are also related to another common application of peer evaluations which is the peer review process that is used to evaluate academic papers for publication. Referee reports play a central role in determining whether a paper is accepted for publication, and success with publishing has a large bearing on academic labor market outcomes. While Abrevaya and Hamermesh [2012] and Blank [1991] do not find a gender disparity in evaluations of paper submissions to an academic journal, there is evidence that papers by female academics face a higher bar for acceptance for publication [Card et al., 2020b, Hengel, 2022]. Recent literature finds that gender disparities in the evaluation of scientific work extend beyond the publication process, including studies by Card et al. [2020a] on peer recognition, Hengel [2019] on citations, Sarsons et al. [2021] on coauthorship and promotions, Arceo-Gomez and Campos-Vazquez [2022] in evaluations of applications for graduate school fellowships, and Chari and Goldsmith-Pinkham [2017] and Hospido and Sanz [2019] on conference submissions.

[3] Some papers have tackled this challenge in other evaluation contexts. For example Card et al. [2020a] and Carrell et al. [2022] study whether editors and referees at academic journals exhibit bias, using citation counts as an objective outcome. Boring [2017] studies bias in teaching evaluations while using a standardized exam score as an objective outcome of teaching/learning performance.

[4] All assignments are typed and submitted through Canvas, which rules out the possibility that graders infer students' gender from their handwriting.

Our paper also contributes to the peer grading literature. While peer grading is a common practice across various educational settings, previous literature on peer grading finds inconclusive evidence on the validity of peer grades[5]. Moreover, there is little to no evidence on gender disparities in peer grading schemes[6]. Inconsistent findings on the validity of peer grading could be due to several factors that differ across educational settings. For example, the degree of clarity in instructions and scoring rubrics has been shown to affect how closely peer grades and instructor-assigned grades correlate. Also, if peer-assigned scores affect students' actual course grades, a competition channel may switch on. Equally plausible, students may be averse to assigning low scores to their classmates[7]. This familiarity effect will also depend on whether students know who their peer grader is, and if they even receive the peer grades at all. We set up our experiment to rule out several of these potential channels, while focusing on the gender differences. The most relevant features of our experiment are that students do not know the identity of their peer evaluator, peer evaluators know the name of the author of the submission that they are assigned to evaluate but are unlikely to know that individual. Peer evaluators also know that peer-assigned grades do not affect the author's course grade. Importantly, our peer evaluators are incentivized to provide an accurate peer assessment. The peer graders' course grades are affected by whether, on average, their peer assessments correlate with the TA's assessments of the same work. This experimental setting allows us to estimate a lower bound for gender disparities by turning off several potential channels for additional bias.

Because we focus on the interaction between the students' and the peer evaluators' genders, our paper is also relevant to the literature on in-group biases. The literature focusing on in-group gender biases provides mixed evidence depending on the context. Boring [2017] and Mengel et al.

[5] Falchikov and Goldfinch [2000] provide a meta-analysis of 56 studies conducted between 1959 and 1999 and find a 69% correlation between student-assigned scores and instructor-assigned scores. Some individual studies, however, find evidence of tighter correlation. Sadler and Good [2006], for instance, analyze self-grading as well as peer grading by middle school students and find a 91-94% correlation with teacher-assigned grades. Luo et al. [2014] analyze data from Massive Open Online Courses (MOOCs) and find that peer grades were fairly similar to teacher-assigned grades on average, and that peer grading improved student learning. Analyzing peer grading in a university setting, Cho et al. [2006] observe that providing clear instructions and rubrics results in closer alignment between peer reviews and teacher-assigned grades.

[6] To our knowledge, there are only two studies on gender bias in peer grading. Sonnert [1995] does not find any gender bias in biologists evaluating their peer scientists' work. In contrast, Langan et al. [2005] find that male undergraduate students favor their male classmates when they evaluate in-class presentations.

[7] Students find it difficult to give negative feedback to classmates, especially friends, because they worry about damaging personal relationships [Lu and Bol, 2007, Topping, 1998].

[2018] show that both female and male students give higher evaluations to male professors. Sarsons et al. [2021] provide experimental evidence for in-group biases in hiring decisions for both males and females. While Bagues et al. [2017] find no evidence that an increase in female evaluators increases the number of chosen female candidates, Bagues and Esteve-Volart [2010] provides evidence that female applicants are less likely to be selected when they are randomly assigned to an evaluation committee with a higher share of women. By analyzing the gender bias in peer evaluations in terms of content and writing skills, our paper also relates to the literature on gender stereotypes where female (male) performance are underestimated (overestimated) in male-stereotyped tasks or fields [Coffman, 2014, Bordalo et al., 2016, 2019, Sarsons et al., 2021, Stoddard and Karpowitz, 2021]. We contribute to this literature by providing compelling evidence that high-achieving men, who will complete performance evaluations in professional settings in the future, exhibit gender bias. This form of bias may contribute to the persistence of gender gaps in professional leadership positions especially in male-dominant fields where evaluators are more likely to be men.

Our results question the validity of peer evaluations broadly. Because we examine an experimental setting in which peer evaluators are incentivized to closely adhere to a clear scoring rubric, these results are especially concerning. In many real-world settings, incentives for objectivity are weak or not present. One reason, which we mention above, is that an objective outcome variable for comparison may not exist. Accordingly, we interpret the gender bias we identify as a lower bound on the corresponding bias that is likely to arise in other institutional settings. The gender biases that we identify suggest that peer evaluations may contribute to gender gaps in hiring, promotion, and wages. For firms, occupations, or industries in which females are in the minority, this observation is particularly concerning.

The rest of the paper proceeds as follows: sections 2 and 3 detail the experimental setting and our data as well as sample selection with descriptive statistics, respectively. Section 4 describes our methodology, defines our variables, and presents our main results. We conclude in Section 5 by discussing our findings.

2 Experimental Setting

There are two primary challenges in identifying and accurately measuring gender biases in peer evaluations. First, evaluators are rarely randomly assigned. Without random assignment, reliable measurements of bias are difficult to obtain. Second, peer evaluations are often undertaken, precisely because an objective measure of performance is unavailable. Accurately attributing an observed disparity to bias requires an objective measure of performance. We overcome these challenges by examining evaluations by randomly assigned peers, in the presence of an objective performance benchmark.

We conducted a peer grading experiment in an introductory economics course at a large, comprehensive research university. Students completed four short essay assignments during the course, which were then evaluated by a randomly matched classmate. Each assignment was based on a clear prompt to which there were objectively correct answers. We compare the peer-assigned grades to official grades of the same assignment submissions and explore whether these two scores systematically differ. Official grades are determined by trained TAs who are randomly assigned to specific students and grade their assignments blindly. TA assignments were re-randomized for each of the four assignments.

The experiment was conducted during the Fall 2018 semester in a large, mostly online introductory macroeconomics course. These undergraduate students came from all undergraduate degree-granting constituent colleges of the university and represented a wide variety of majors. Enrolled students were 44.95% male and 55.05% female.

All enrolled students were required to complete four short essay assignments. These essay assignments asked students to answer specific questions about an economic graph. There was a single objectively correct answer to each question. No outside research was required, and neither subjective analysis nor students' own opinions were solicited. These questions were the types of questions that an instructor would typically include as a free-response question on an exam in a smaller course. Students were told that their submission should be composed in "essay form", and that while economics content would play a much larger role in determining their grade, writing quality would also play a role. They were also told that there was no minimum or maximum required length, but that a strong answer could be provided in approximately 150 words[8].

Assignments were submitted into an electronic course management system (Canvas). Each submission was graded blindly by a randomly assigned trained TA[9], who assigned the official score for inclusion in the student's final course grade calculation, and by a randomly assigned (non-blind) peer grader for evaluation. Both blind TAs and non-blind peer graders evaluated the assignments using a common scoring rubric. Rubrics contained two types of questions: between seven and eleven economic content questions, each of which had an objectively correct answer, and three writing quality questions. Each question on the rubric was awarded a numeric score between zero and ten. Partial credit was available for economic content questions if, for example, a student provided the correct answer but used incorrect units of measurement. Writing questions were scored on the same zero-to-ten scale. The grading procedure generated a content subscore (defined as the percentage of possible content points earned), a writing subscore (defined as the percentage of possible writing points earned), and an overall assignment grade (defined as the percentage of total possible points earned). Peer graders had one week to complete the peer review. Official TA grades were released after one week. Neither group had access to the other group's scores.

When completing their evaluations, both TAs and peer graders had access to the same information and scoring rubrics. The only difference was that peer graders could view the submission author's first and last name and possibly a small picture of the student's face. The TAs graded blindly; they did not have access to the names or the pictures of the authors. Importantly, this experiment took place in a very large class. Most of students did not know each other. The majority of peer graders did not know the gender of the author. Instead, they could infer the author's gender from the name or possibly look at the picture of the student.[10]. Rather than using the students' actual gender, we use the gender predicted by their names. Using genderize.io, we predict the gender probabilities for all first names. To be conservative, we used the following definition: If gender is predicted as female with more than 90% probability, we define the name as female-sounding. Similarly, if gender is predicted as male with more than 90% probability, we define the name as male sounding. The remaining names are treated as ambiguous, and we

[8] A sample assignment prompt is provided in Figure 1

[9] There were 5 male and 4 female TAs. All TAs were economics graduate students.

[10] Some students did not upload pictures into the system, and even when they did, these pictures were very small (i.e., approximately the size of a dime).

assign "unknown gender." We complete robustness checks by performing the analysis with a 70% probability threshold as well as using the probability of being female as an alternative measure.

Only TA-assigned grades were used to calculate students' final course grades. Each assignment accounted for 2% of the author's final course grade. Peer grades only affected the final course grade of the peer grader. Peer graders received an overall peer grading score (across all four peer reviews) that accounted for 4% of their final course grade. They did not receive any feedback on the quality of their peer reviews before the end of the course. That is, they could not adjust their approach assignment-to-assignment in response to feedback. Peer graders were told that they must complete a peer review and that it must "more or less" match the TA's evaluations to receive credit. If the reviews were too dissimilar from the TA's grades, they would not receive credit for completing the peer review. The clear incentive was to match the TAs' evaluations. If students had prior beliefs about the strictness of TAs' grades, they should have incorporated those beliefs when completing their peer reviews. These instructions are consistent with the analysis that follows in the paper, which focuses on the correlation between TA-assigned and peer review scores.

It is important to highlight the most relevant features of the experimental setting, as it relates to accurately identifying and measuring the gender biases in peer evaluations.

In our setup, peer-assigned grades do not affect the author's final course grade. Studying settings in which peer evaluations affect someone else's grade (or any other outcome or determinant of well-being) would provide valuable insights, because that is what often happens in "real-world" settings. However, if the peer grades were included in final course grades, rather than TA-assigned grades, it would become challenging to isolate individual channels that may be at play. In our setup, for example, we do not expect to observe any competition or familiarity effects that could generate systematic deviations from the TA-assigned grades, which could also differ by gender.

Last but not least, peer graders' final course grades are affected by their careful and thoughtful completion of peer reviews. This incentivizes them to be as precise as possible. The discrimination literature routinely shows that the observed discrimination tends to be lower or diminish when there is explicit or implicit monitoring (See for example, Parsons et al. [2011]). Moreover, because of these clear incentives, we interpret the gender biases that we identify as "lower bounds."

3 Sample Selection and Summary Statistics

The data are generated by the randomly assigned blind teaching assistants' expert evaluations and randomly assigned non-blind classmates' peer reviews of four short writing assignments. After the submission deadline, each student's assignment was randomly assigned to a specific TA's grading queue and randomly assigned to one of the student's classmates for peer review. The peer review assignments were made randomly by the course management system, and the matching process was based solely on the pool of students who submitted the assignment by the deadline. If a student did not submit their assignment by the deadline, they received a zero for their own assignment and were not assigned a peer review, which implied a zero on that peer review exercise as well. Each student that submitted an assignment by the deadline was assigned a peer review, and each peer reviewer was only assigned one submission to review.

In our data, we observe all students who were ever enrolled in the course in Fall 2018. On the other hand, in our peer grading analysis, inclusion in the sample is conditional on submitting the assignment correctly by the deadline. Students who failed to submit a particular assignment by the submission deadline as instructed are not included in the data for that assignment. These students will, however, appear in the data for other assignments. Also, we have missing TA-assigned or peer-assigned grades if students fail to follow submission instructions and/or if a peer grader does not complete their assigned peer review. Ultimately, we work with the sample of homework assignments that are submitted correctly, and for which the peer reviewer completed their review.

In panel A of Table 1, we report the descriptive statistics on the full sample of 975 students for 4 assignments leading to 3900 observations over the semester. Approximately 24% of assignments were either not submitted, submitted late, or submitted incorrectly. Among the assignments that were submitted on-time and correctly, 2% of assigned peer reviewers did not complete the review. In terms of incomplete peer grading, we do not find any significant differences between men and women. On the other hand, male students were more likely to skip an assignment or submit late/incorrectly, which is the only apparent significant gender difference in terms of the dropped observations.

Descriptive statistics of performance measures are reported for the full sample in panel B of

Table 1. We report the same statistics for the restricted sample used in our analysis in panel C of Table 1. Female students perform better than male students on all assessments in the course, except for the three high-stakes exams. This observation is consistent with previous findings on gender differences in performance under pressure and when stakes are high. There are small differences between the restricted and full sample. The restricted sample is slightly more positively selected in terms of performance in assessments, but the gender differences in TA-assigned scores and peer scores, and gender composition appear similar.

The summary statistics in panel C of Table 1 reveal that there is a small performance gap in favor of female students for TA-assigned scores. This gap is largely driven by content subscores. The difference between male and female students' TA-assigned writing subscores is comparably very small and just barely significant. For the peer-assigned grades, the gender gap is larger, but it follows the same pattern in content and writing subscores. The peer-assigned scores appear to be lower than the TA-assigned scores both for female and male students.

We check for balance of characteristics between the assignments graded by female and male peers. In Table 2, we show the summary statistics of peer assigned scores by the gender of the peer grader. It seems that peer graders were not more or less likely to be assigned to a student with a certain gender in our randomized peer assignment. Also, female and male peer graders seem to be assigned to students who performed similarly in terms of blind TA-assigned scores and other overall course performance measures. This table also shows that female peer graders, on average, give lower scores to the homework assignments they grade than the male peer graders. This gap seems to be the larger in content subscores. Based on the summary statistics, it appears that peer graders are tougher graders than the TAs, and that female peers deviate even more *negatively* than male peer graders. In the analysis below, we will investigate these differences more carefully.

Lastly, we also check for balance of characteristics between the assignments blindly graded by female and male teaching assistants. Table 3 shows the summary statistics by the gender of the teaching assistants. These statistics suggest no significant differences in terms of performance measures. We do not find any statistically significant difference in TA-assigned overall scores or content subscores. However, it appears that male TAs assign slightly higher writing subscores than their female counterparts.

To check whether the validity of peer grading is similar across the distribution of scores, we first plot the cumulative distribution functions for overall scores, content subscores, and writing subscores, separately for TA-assigned scores and peer scores. These CDFs are depicted in Figure 2. These figures reveal that the peer content subscores are lower than TA-assigned content subscores across the distribution, and these differences are slightly larger at the higher end of the score distributions.

Table 2 shows that assignments graded by female peer graders have lower scores on average, and that this difference is driven by content subscores. In order to observe whether this observation holds across the distribution of scores, we plot the cumulative distribution functions of each subscore assigned by peer graders, separately for male and female peer graders. These CDFs are presented in Figure 3. These figures reveal that the differences are larger in the content subscores, and that content subscores by male peer graders stochastically dominate content subscores by female peer graders.

Finally, we consider the deviation of peer grades from the blind TA-assigned grades. Figure 4 shows the average deviations by gender pairs, where author's gender is predicted by their first name. This figure shows that peer graders deviate negatively, on average. These negative deviations are similar across different gender pairs, except when male peer graders are randomly assigned to students without female sounding names. The second graph in this figure further distinguishes the gender pair. Each author's gender is categorized by a female-sounding name, male sounding name, or ambiguous name[11]. This graph is consistent with the previous one, again suggesting similar negative deviations for all gender pairs, except when male peer graders are randomly assigned to authors with male-sounding names or ambiguous names. Figure 5 shows the deviations between TA-assigned and peer scores, by gender pairs, for the content and writing subscores. It suggests that the differences are driven by the deviations in content subscores. The deviations in writing subscores do not appear to be statistically different from zero for any gender pair. In the next section, we analyze these differences in more detail.

[11] If a name is predicted to be a female or male name with a 90% or higher probability, we define it as a female or male sounding name. All other names are defined as ambiguous.

4 Methodology and Results

The empirical strategy employs a fixed effects model. There are four short writing assignments in the semester, and blind TA and peer grader assignments are re-randomized each time. Accordingly, the same grader evaluates different students across the four assignments, and each student is graded by different graders. We use peer grader fixed effects, which controls for unobserved grader characteristics. We compare peer-assigned scores when the peer grader randomly receives a submission with a female- versus a male-sounding name, while controlling for the (blind) TA-assigned grade. Similarly, we include student-fixed effects to control for unobserved student characteristics. We compare the peer-assigned grades the same student is randomly assigned to a female or male grader, again controlling for the (blind) TA-assigned grade. We also have the opportunity to compare the peer-assigned grades when the same student is randomly assigned to a high-performing or a low-performing peer grader.

Following Boring [2017], we first use peer grader fixed effects to compare how the same (male or female) peer grader evaluates a submission by a randomly assigned student with a female-sounding name with a randomly assigned student without a female-sounding name.

$$\text{Peer Score}_{ijt} = \alpha + \beta_1 \text{Stu Peer M}_{ij} + \beta_2 \text{Stu Peer F}_{ij} + \beta_3 X'_{it} + \beta_4 \text{Z}_{jt} + \mu_j + \epsilon_{ijt} \tag{1}$$

where Peer Score$_{ijt}$ is the score student i receives from peer grader j on assignment t. The two main variables of interest, when considering the effect of gender pairs, are Stu Peer M$_{ij}$, which is a dummy variable that takes the value 1 if the peer score is given by a male peer grader to a student without a female-sounding name, and Stu Peer F$_{ij}$, which is a dummy variable that takes the value 1 if the peer score is given by a female peer grader to a student with a female-sounding name. X'_{it} includes the TA-assigned score that student i receives on assignment t, as well as the TA's gender. Z'_{jt} includes peer grader j's own TA-assigned scores on assignment t, and μ_j represents the peer grader fixed effects.

Similarly, in the student fixed effects specification, we are able to compare how the same student is graded by a randomly assigned male peer grader versus a randomly assigned female peer grader:

$$\text{Peer Score}_{ijt} = \alpha + \beta_1 \text{Stu Peer M}_{ij} + \beta_2 \text{Stu Peer F}_{ij} + \beta_3 X'_{it} + \beta_4 \text{Z}_{jt} + \nu_j + \epsilon_{ijt} \qquad (2)$$

These fixed effects models enable us to control for three important unobservable student and peer grader characteristics that may influence the peer assigned scores. These unobserved characteristics are students' knowledge, their ability to demonstrate knowledge in an essay, and their grading styles as peer reviewers. While controlling for blind TA-assigned scores, we can analyze whether male peer graders are biased against female students, whether female peer graders are biased against male students, whether peer graders in general are biased against female students, and whether there is a difference in how biased male and female peer graders are.

We begin with the analysis of the content subscores. Recall that these content subscores are based on questions with objectively correct answers. The scoring rubric clearly establishes the points that should be awarded for correct or partially correct answers. Table 4 shows the results from the peer grader fixed effects model in the first two columns and the student fixed effects model in the last two columns. Standard errors are clustered at student and peer grader level. The results from the peer grader fixed effects models show that male peer graders give higher content scores to students without female-sounding names, relative to assignments with female-sounding names. The results in columns 1 and 2 suggest that male peer graders award students without female-sounding names content subscores that are approximately 2.1 points higher, compared to what they award students with female-sounding names. The student fixed effects models in columns 3 and 4 show that students without female-sounding names receive content subscores that are approximately 2.3 points higher with male peer graders than with female peer graders.

Both of these specifications in equation 1 and equation 2 give the same results as if each of them were estimated as two separate specifications: one in which Stu Peer $\text{F}_i j$ is kept and Stu Peer $\text{M}_i j$ is replaced by a student's predicted gender dummy variable, and another in which Stu Peer $\text{M}_i j$ is kept and the Stu Peer $\text{F}_i j$ variable is replaced by a student's predicted gender dummy variable. We provide the results on these separate estimations both for peer grader and student fixed effects regressions in the table 7. Using equation 1, the result on Stu Peer $\text{M}_i j$ directly shows how a same male peer grader grades a student without a female sounding name compared to a student with a female sounding name. The result on Stu Peer F_ij directly shows how a same female peer grader grades a student with a female sounding name compared to a student without a female sounding name.

According to columns 1 and 2 of Table 4, female peer graders award similar content subscores to students with female-sounding names, compared to students without female-sounding names. The estimated difference is around 0.2 points, suggesting a possible small and statistically insignificant bias in favor of female-sounding names. Similarly, students with female-sounding names tend to receive similar grades from male and female peer graders, according to the specifications in Columns 3 and 4. Column 3 and 4 identify small and statistically insignificant estimates that suggest students with female-sounding names may receive lower grades if they are randomly assigned a female peer grader.

Previously, we provided balance checks, both by the TA's gender and peer grader's gender. The key balance issue in this analysis is whether male and female peer graders are randomly assigned similar pools of students with and without female-sounding names. We could be concerned, for instance, if male peer graders only receive assignments from poorly qualified students with female-sounding names and highly qualified students without female-sounding names, whereas female peer graders receive assignments with similar qualifications from students with and without female-sounding names. This possibility is not necessarily ruled out by our initial balance checks. To address this potential concern, we also run a regression of blind TA-assigned scores on the gender pairs, with student fixed effects. Table 5 shows that there is no statistically significant difference in TA-assigned scores across different gender pairs.

Next, we conduct a similar analysis of the writing subscores, which are more susceptible to subjective judgment, because there is no single "correct" way to compose an essay response. Table 6 presents the results for the writing subscores using the same specifications and format as Table 4. The results from the peer grader fixed effects models in Columns 1 and 2 show that male peer graders award lower writing subscores to students without female-sounding names, compared to how they evaluate assignments with female-sounding names. The estimates suggest that male peer graders assign approximately 1.5 fewer points in the writing subscore to students without female-sounding names, compared to students with female-sounding names. The student fixed effects regressions in Columns 3 and 4 show that students without female-sounding names receive approximately 0.3 fewer points in the writing subscore when they are randomly assigned a male peer grader, compared to being randomly assigned a female peer grader. This estimate is small and not statistically significant. These results suggest that the advantage given by male graders to students without female-sounding names with content subscores is reversed in favor of female-sounding names with writing subscores. This observation is consistent with stereotypes that women are better at writing.

While male peer graders assign higher writing subscores to assignments with female-sounding names, this effect is dominated by the opposiste effect on the content subscores. The magnitude of male peer graders' bias in favor of assignments without female-sounding names in the content subscores (approximately 2.1 points) exceeds that of their bias in favor of assignments with female-sounding names in the writing subscores (approximately 1.5 points). These are independent estimates, which implies that the relative size of these biases is not driven by the outsized importance of content subscores in overall grade calculations. However, that weighting does reinforce the difference.

Female graders award similar writing subscores to students with female-sounding names, compared to students without female sounding names. The estimates in Columns 1 and 2 of Table 6 suggest a small, approximately 0.6-point bias against female-sounding names, but this bias is statistically insignificant. Similarly, Columns 3 and 4 suggest that students with female-sounding names receive slightly lower writing subscores from female peer graders, compared to male peer graders. These results are also statistically insignificant.

Table 4 and 6 also provide insights into the general validity of peer grading. We find that content subscores demonstrate considerably more validity than writing subscores. In Table 4, the coefficient for the content subscore given blindly by TAs is approximately 0.79 when measured using peer grader fixed effects and 0.69 when measured using student fixed effects. This result can be interpreted as saying that a 1-point increase in the TA-assigned content subscore leads to a 0.69-0.79-point increase in the peer-assigned content subscore. Unsurprisingly, we observe much more severe validity concerns in the writing subscores. In Table 6, the coefficient for the blind

TA writing subscore is approximately 0.11 when measured using peer grader fixed effects and 0.08 when measured using student fixed effects. Moreover, an 8-11% correlation is especially low and concerning when establishing validity.

Finally, in Table 4 and 6, we observe the effect of own performance on peer graders scoring behavior. In the fourth column of both tables, in which we apply student fixed effects, the coefficients of graders own content subscore suggest that when a same student is randomly assigned to a peer grader whose content subscore is higher, they receive lower peer-assigned scores, conditional on their TA-assigned scores. When the randomly-assigned peer reviewer's own content subscore increases by one point, a student receives 0.37 fewer points in the content subscore and 0.4 fewer points in the writing subscore from the peer reviewer. This result suggests that higher performing peer reviewers assign lower peer review scores.

In all of these regressions, we define female-sounding names as names that are predicted to be a female name with more than 90% probability. In order to check for robustness of our results to this definition, we complete robustness checks by performing the analysis with a 70% probability threshold (in Table 9 and Table 11 as well as using the probability of being female as an alternative measure in Table 8 and Table 10 where we do not find significantly different results.

In this section, we present several results related to the validity of peer grading and gender biases. First, we investigate the general validity of peer grading. We find that peer graders assign lower scores than TAs assign to the same submissions[12]. When we decompose this effect by focusing separately on content and writing subscores, we identify stronger validity in the content subscores and much weaker validity in the writing subscores. Next, we find that male peer graders assign higher scores to students without female-sounding names. Similarly, students without female-sounding names receive high peer scores when they are randomly assigned to a male peer reviewer. These effects are driven by the content subscores. We do not find evidence that female peer graders evaluate students with or without female-sounding names differently, nor that a student with a female-sounding name receives higher or lower peer scores when randomly assigned to a female peer reviewer. Finally, we observe that higher performing peer graders assign lower peer scores.

These findings are particularly compelling, because they were obtained in an experimental

[12] Students do not know their TA graders and all grade contests are sent to the instructor. Given that TAs do not deal with grade contests, we do not expect the TAs to give higher grades to avoid student complaints.

setting in which peer graders were incentivized to match a clear scoring rubric. For this reason, we view the observed gender bias as a lower bound estimate. The results have strong implications for real-world settings, especially when female students and workers are the minority in a class, field, industry, or occupation and are likely to face male peer evaluators.

5 Conclusion

We conduct a peer evaluation experiment using peer grading in a large introductory economics course at a flagship public research university. Students complete short writing assignments, for which there is an objectively correct answer, and each submission is evaluated blindly by a randomly assigned trained graduate teaching assistant and non-blindly by a randomly-assigned classmate. Overall grades are calculated as a weighted combination of content and writing subscores, where the latter is expected to have more room for subjective evaluation. The course management system (Canvas) randomly assigns each submission to one of the author's classmates for peer review. The peer reviewer then evaluates their classmate's submission using the same rubric as the teaching assistants, and they are told that they should "more or less" match the TA-assigned grades in order to receive credit.

We compare (blind) TA-assigned grades and (non-blind) peer-assigned grades to evaluate the general validity of peer grading in university courses and to identify whether gender biases may explain any observed divergence between these two sets of scores. We assert that peer grading is valid if peer grades match the grades assigned by trained TAs. Beyond simply determining whether these grades match, we also examine whether any observed differences exhibit an identifiable gender bias.

Our findings can be summarized in four pillars. First, we find that expert and peer grades differ systematically. Validity concerns are more severe when evaluating writing quality than when evaluating answers to specific economics content questions with objectively correct answers. One might be concerned that TAs may be more prone to inflating grades if they wish to avoid complaints and contested scores. This could generate the observed deviations between TA and peer grades. However, in this experimental setting, complaints and contested scores were not handled by grading

TAs, but instead by the course instructor. The TAs were simply instructed to evaluate the work blindly and assign a grade using a clear scoring rubric. There were no apparent incentives to inflate or deflate scores by randomly assigned blind teaching assistants.

Second, we observe gender bias in content subscores. Male peer graders assign 2.02-point higher scores, on average, to authors without a female-sounding name than they do to authors with a female-sounding name. Similarly, students without a female-sounding name receive 2.50-point higher scores, on average, when they are randomly assigned to a male peer grader. While these are not small deviations in magnitude, they are also statistically significant and can be interpreted as lower bound estimates. Because peer graders were incentivized to match the TA-assigned scores, the observed deviation is potentially smaller than it would be without "monitoring." Monitoring has been shown to reduce biases and result in more accurate peer evaluations.

Observing clear gender bias in the content subscores but not in the writing subscores is suggestive evidence that the bias stems from statistical discrimination, as opposed to taste-based discrimination. Taste-based discrimination would be expected to appear in both subscores. Because it is only observed in one – and more interestingly, not the more subjective writing subscores – we suspect that the observed bias is consistent with statistical discrimination.

Female peer graders assign lower scores than male peer graders on average. By using blind TA-assigned scores as an objective benchmark, and analyzing peer scores by gender pairs, we establish the specific channel through which this disparity arises. It is not that female peer graders are tougher graders, *per se*. Rather, male peer graders give a positive boost to their male classmates. Peer scores assigned by female students, in general, and peer scores assigned by male students to female students are quite similar.

Third, we also find that high performing peer graders are tougher peer graders. We also considered whether course performance and understanding of the material might differ by gender, which could explain differences in peer grading practices. However, conditioning on the peer grader's own score did not affect the results. Female students perform better on these assignments, but that is not driving the difference in scores assigned by female and male peer graders.

While we observe some evidence of gender bias, it is possible that the experimental setting itself combats such a bias. Because peer graders are monitored and even incentivized, they may demonstrate less bias than they would in a “real world” setting without monitoring and/or incentives. That is, it may be the case that peer graders possess gender-based biases, but they did not act on those biases because they knew that they were being observed. Moreover, in an experimental setting, students with female sounding names were equally likely to be assigned to female or male peer graders. In a “real world” setting in which male evaluators favor males, female students and workers in male-dominant fields will be more likely to be receiving lower grades or performance evaluations than their male counterparts, precisely because peer evaluators are more likely to be male.

References

Jason Abrevaya and Daniel S. Hamermesh. Charity and favoritism in the field: Are female economists nicer (to each other)? *The Review of Economics and Statistics*, 94(1):202–207, 2012. ISSN 00346535, 15309142. URL http://www.jstor.org/stable/41349169.

Eva O. Arceo-Gomez and Raymundo M. Campos-Vazquez. Gender bias in evaluation processes. *Economics of Education Review*, 89:102272, 2022. ISSN 0272-7757. doi: https://doi.org/10.1016/j.econedurev.2022.102272. URL https://www.sciencedirect.com/science/article/pii/S0272775722000486.

Manuel Bagues, Mauro Sylos-Labini, and Natalia Zinovyeva. Does the gender composition of scientific committees matter? *American Economic Review*, 107(4):1207–38, April 2017. doi: 10.1257/aer.20151211. URL https://www.aeaweb.org/articles?id=10.1257/aer.20151211.

Manuel F. Bagues and Berta Esteve-Volart. Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment. *The Review of Economic Studies*, 77(4):1301–1328, 10 2010. ISSN 0034-6527. doi: 10.1111/j.1467-937X.2009.00601.x. URL https://doi.org/10.1111/j.1467-937X.2009.00601.x.

Lori Beaman, Niall Keleher, and Jeremy Magruder. Do job networks disadvantage women? evidence from a recruitment experiment in malawi. *Journal of Labor Economics*, 36(1):121–157, 2018. doi: 10.1086/693869. URL https://doi.org/10.1086/693869.

Rebecca M. Blank. The effects of double-blind versus single-blind reviewing: Experimental evidence from the american economic review. *The American Economic Review*, 81(5):1041–1067, 1991. ISSN 00028282. URL http://www.jstor.org/stable/2006906.

Pedro Bordalo, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. Stereotypes*. *The Quarterly Journal of Economics*, 131(4):1753–1794, 07 2016. ISSN 0033-5533. doi: 10.1093/qje/qjw029. URL https://doi.org/10.1093/qje/qjw029.

Pedro Bordalo, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. Beliefs about gender.

American Economic Review, 109(3):739–73, March 2019. doi: 10.1257/aer.20170007. URL https://www.aeaweb.org/articles?id=10.1257/aer.20170007.

Anne Boring. Gender biases in student evaluations of teaching. *Journal of Public Economics*, 145:27–41, 2017. ISSN 0047-2727. doi: https://doi.org/10.1016/j.jpubeco.2016.11.006. URL https://www.sciencedirect.com/science/article/pii/S0047272716301591.

David Card, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri. Gender differences in peer recognition by economists. *Working Paper*, 2020a.

David Card, Stefano DellaVigna, and Nagore Iriberri. Are referees and editors in economics gender neutral? *The Quarterly Journal of Economics*, 135(1):269–327, 2020b. doi: 10.1093/qje/qjz035. URL https://doi.org/10.1093/qje/qjz035.

Scott E Carrell, David N Figlio, and Lester R Lusher. Clubs and networks in economics reviewing. Working Paper 29631, National Bureau of Economic Research, January 2022. URL http://www.nber.org/papers/w29631.

Paola Cecchi-Dimeglio. How gender bias corrupts performance reviews, and what to do about it. *Harvard Business Review*, 2017. URL https://hbr.org/2017/04/how-gender-bias-corrupts-performance-reviews-and-what-to-do-about-it.

Anusha Chari and Paul Goldsmith-Pinkham. Gender representation in economics across topics and time: Evidence from the nber summer institute. Working Paper 23953, National Bureau of Economic Research, October 2017. URL http://www.nber.org/papers/w23953.

Kwangsu Cho, Christian D. Schunn, and Roy W. Wilson. Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, 98(4):891–901, 2006. doi: 10.1037/0022-0663.98.4.891. URL https://doi.org/10.1037/0022-0663.98.4.891.

Katherine Baldiga Coffman. Evidence on Self-Stereotyping and the Contribution of Ideas *. *The Quarterly Journal of Economics*, 129(4):1625–1660, 09 2014. ISSN 0033-5533. doi: 10.1093/qje/qju023. URL https://doi.org/10.1093/qje/qju023.

Shelley J. Correll and Caroline Simard. Vague feedback is holding women back. *Harvard Business Review*, 2016. URL https://hbr.org/2016/04/research-vague-feedback-is-holding-women-back.

Alessandro Di Fiore and Souza Marcio. Are peer reviews the future of performance evaluations? *Harvard Business Review*, 2021. URL https://hbr.org/2021/01/are-peer-reviews-the-future-of-performance-evaluations.

Nancy Falchikov and Judy Goldfinch. Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, 70(3):287–322, 2000.

Claudia Goldin and Cecilia Rouse. Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, 90(4):715–741, September 2000. doi: 10.1257/aer.90.4.715. URL https://www.aeaweb.org/articles?id=10.1257/aer.90.4.715.

Erin Hengel. Gender differences in citations at top economics journals. *Working Paper*, 2019.

Erin Hengel. Are Women Held to Higher Standards? Evidence from Peer Review*. *The Economic Journal*, 05 2022. ISSN 0013-0133. doi: 10.1093/ej/ueac032. URL https://doi.org/10.1093/ej/ueac032. ueac032.

Laura Hospido and Carlos Sanz. Gender gaps in the evaluation of research: Evidence from submissions to economics conferences. *Center for Economic and Policy Research Discussion Paper*, 2019. URL http://ftp.iza.org/dp12494.pdf.

A. Mark Langan, C. Philip Wheater, Emma M. Shaw, Ben J. Haines, W. Rod Cullen, Jennefer C. Boyle, David Penney, Johan A. Oldekop, Carl Ashcroft, Les Lockey, and Richard F. Preziosi. Peer assessment of oral presentations: effects of student gender, university affiliation and participation in the development of assessment criteria. *Assessment & Evaluation in Higher Education*, 30(1):21–34, 2005. doi: 10.1080/0260293042003243878. URL https://doi.org/10.1080/0260293042003243878.

LeanIn.Org and McKinsey & Company. Women in the work place 2022. 2022.

Ruiling Lu and Linda Bol. A comparison of anonymous versus identifiable e-peer review on college student writing performance and the extent of critical feedback. *Journal of Interactive Online Learning*, 6(2):100–115, 2007. ISSN 1541-4914.

Heng Luo, Anthony C. Robinson, and Jae Young Park. Peer grading in a mooc: Reliability, validity, and perceived effects. *Online Learning Journal*, 18(2), 2014. ISSN 2472-5730. doi: 10.24059/olj.v18i2.429.

Friederike Mengel, Jan Sauermann, and Ulf Zölitz. Gender Bias in Teaching Evaluations. *Journal of the European Economic Association*, 17(2):535–566, 02 2018. ISSN 1542-4766. doi: 10.1093/jeea/jvx057. URL `https://doi.org/10.1093/jeea/jvx057`.

K. L. Milkman, M. Akinola, and D Chugh. What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *The Journal of Applied Psychology*, 100(6):1678–1712, 02 2015. doi: 10.1037/apl0000022. URL `https://doi.org/10.1037/apl0000022`.

Christopher A. Parsons, Johan Sulaeman, Michael C. Yates, and Daniel S. Hamermesh. Strike three: Discrimination, incentives, and evaluation. *The American Economic Review*, 101(4): 1410–1435, 2011. ISSN 00028282. URL `http://www.jstor.org/stable/23045903`.

Natasha Quadlin. The mark of a woman's record: Gender and academic performance in hiring. *American Sociological Review*, 83(2):331–360, 2018. doi: 10.1177/0003122418762291. URL `https://doi.org/10.1177/0003122418762291`.

Philip M. Sadler and Eddie Good. The impact of self- and peer-grading on student learning. *Educational Assessment*, 11(1):1–31, 2006.

Heather Sarsons. Referrals interpreting signals in the labor market: Evidence from medical referrals. *Working Paper*, 2017.

Heather Sarsons, Klarita Gërxhani, Ernesto Reuben, and Arthur Schram. Gender differences in recognition for group work. *Journal of Political Economy*, 129(1):101–147, 2021. doi: 10.1086/711401. URL `https://doi.org/10.1086/711401`.

Gerhard Sonnert. What makes a good scientist?: Determinants of peer evaluation among biologists. *Social Studies of Science*, 25(1):35–55, 1995. doi: 10.1177/030631295025001003. URL https://doi.org/10.1177/030631295025001003.

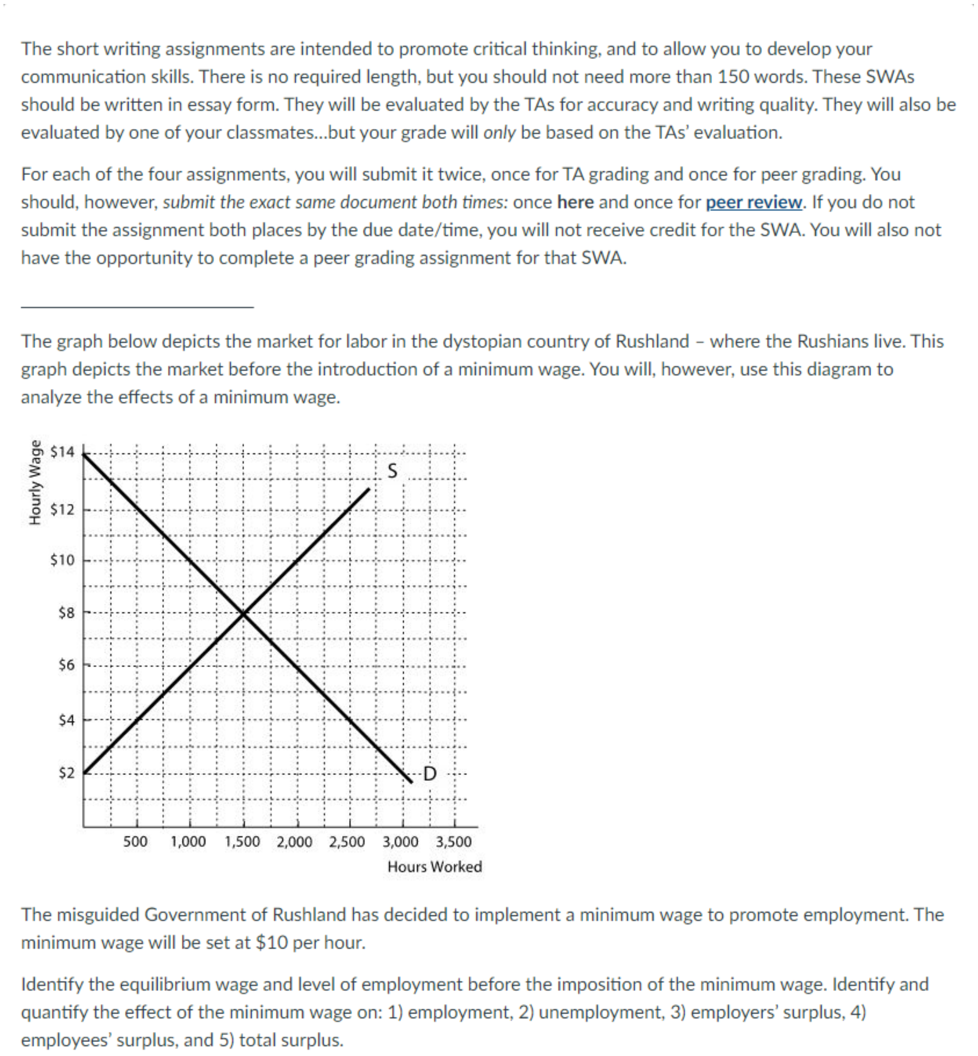
Olga Stoddard and Preece Jessica Karpowitz, Christopher F. Strength in numbers: A field experiment in gender, influence, and group dynamic. *mimeo*, 2021.

Keith Topping. Peer assessment between students in colleges and universities. *Review of Educational Research*, 68(3):249–276, 1998. ISSN 00346543, 19351046. URL http://www.jstor.org/stable/1170598.

Dan Zeltzer. Gender homophily in referral networks: Consequences for the medicare physician earnings gap. *American Economic Journal: Applied Economics*, 12(2):169–97, April 2020. doi: 10.1257/app.20180201. URL https://www.aeaweb.org/articles?id=10.1257/app.20180201.

6 Figures and Tables

Figure 1: Sample Prompt for a Short Writing Assignment



The short writing assignments are intended to promote critical thinking, and to allow you to develop your communication skills. There is no required length, but you should not need more than 150 words. These SWAs should be written in essay form. They will be evaluated by the TAs for accuracy and writing quality. They will also be evaluated by one of your classmates...but your grade will *only* be based on the TAs' evaluation.

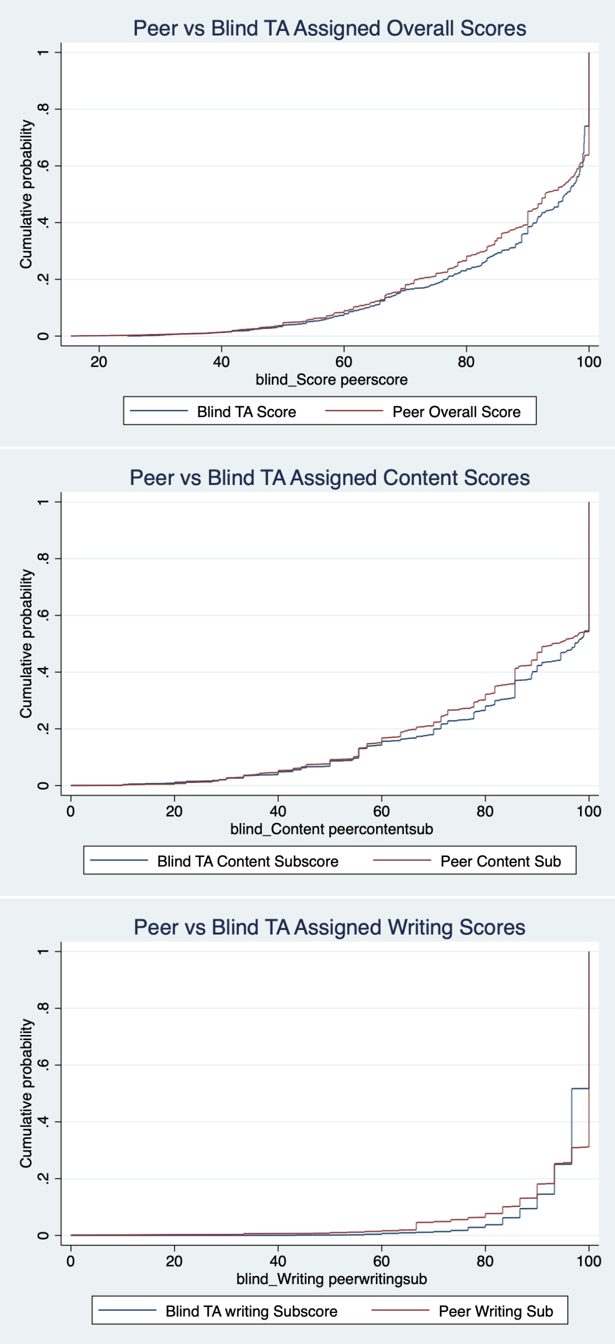
For each of the four assignments, you will submit it twice, once for TA grading and once for peer grading. You should, however, *submit the exact same document both times:* once **here** and once for **peer review**. If you do not submit the assignment both places by the due date/time, you will not receive credit for the SWA. You will also not have the opportunity to complete a peer grading assignment for that SWA.

The graph below depicts the market for labor in the dystopian country of Rushland – where the Rushians live. This graph depicts the market before the introduction of a minimum wage. You will, however, use this diagram to analyze the effects of a minimum wage.

The misguided Government of Rushland has decided to implement a minimum wage to promote employment. The minimum wage will be set at $10 per hour.

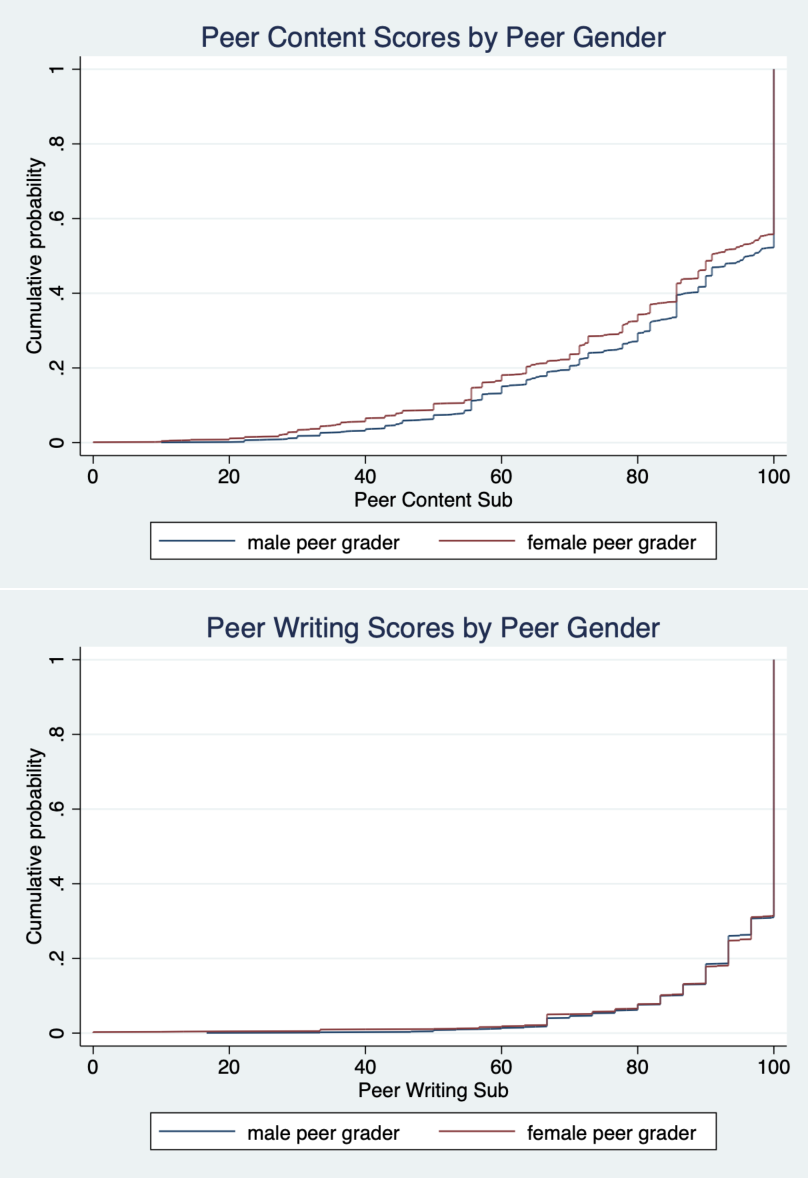
Identify the equilibrium wage and level of employment before the imposition of the minimum wage. Identify and quantify the effect of the minimum wage on: 1) employment, 2) unemployment, 3) employers' surplus, 4) employees' surplus, and 5) total surplus.

Figure 2: CDF of TA-assigned and Peer-assigned Grades



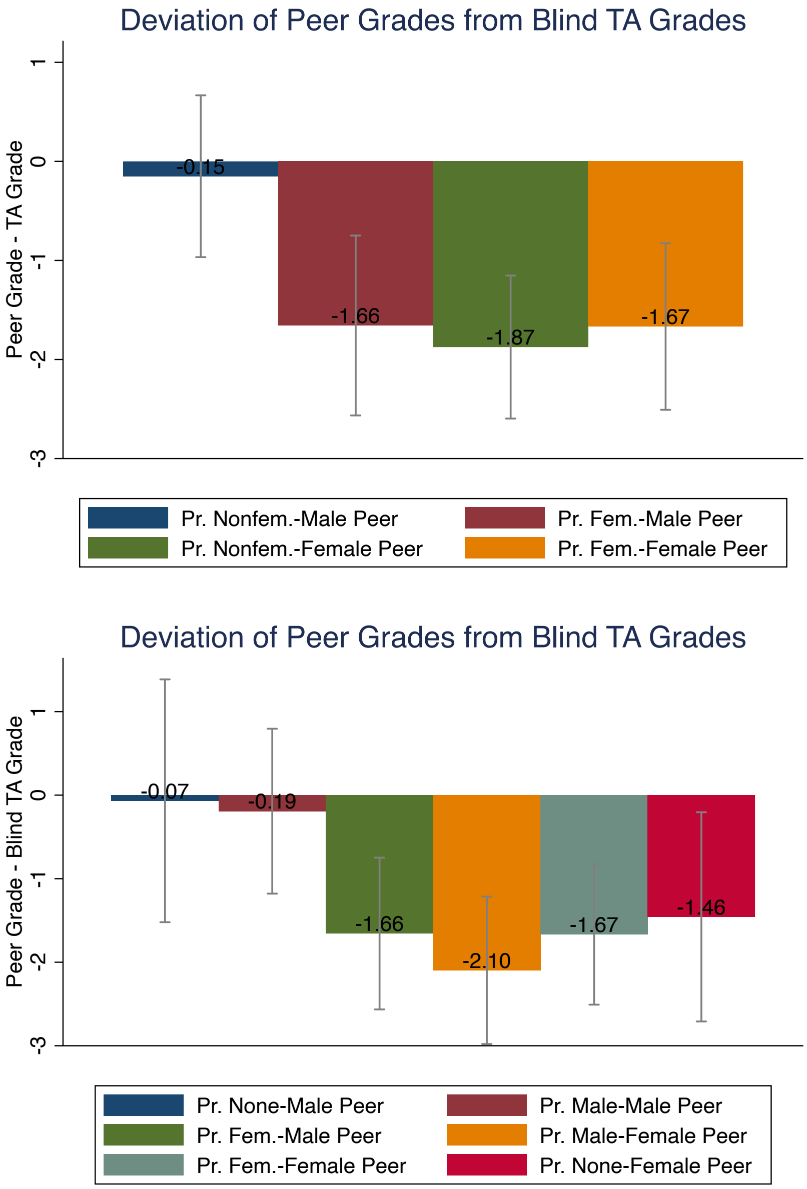
Note: The figures present the cumulative distribution functions for overall scores, content subscores, and writing subscores, separately for blind TA-assigned scores and non-blind peer scores.

Figure 3: CDF of Subscores Assigned by Peers by Peer Grader's Gender



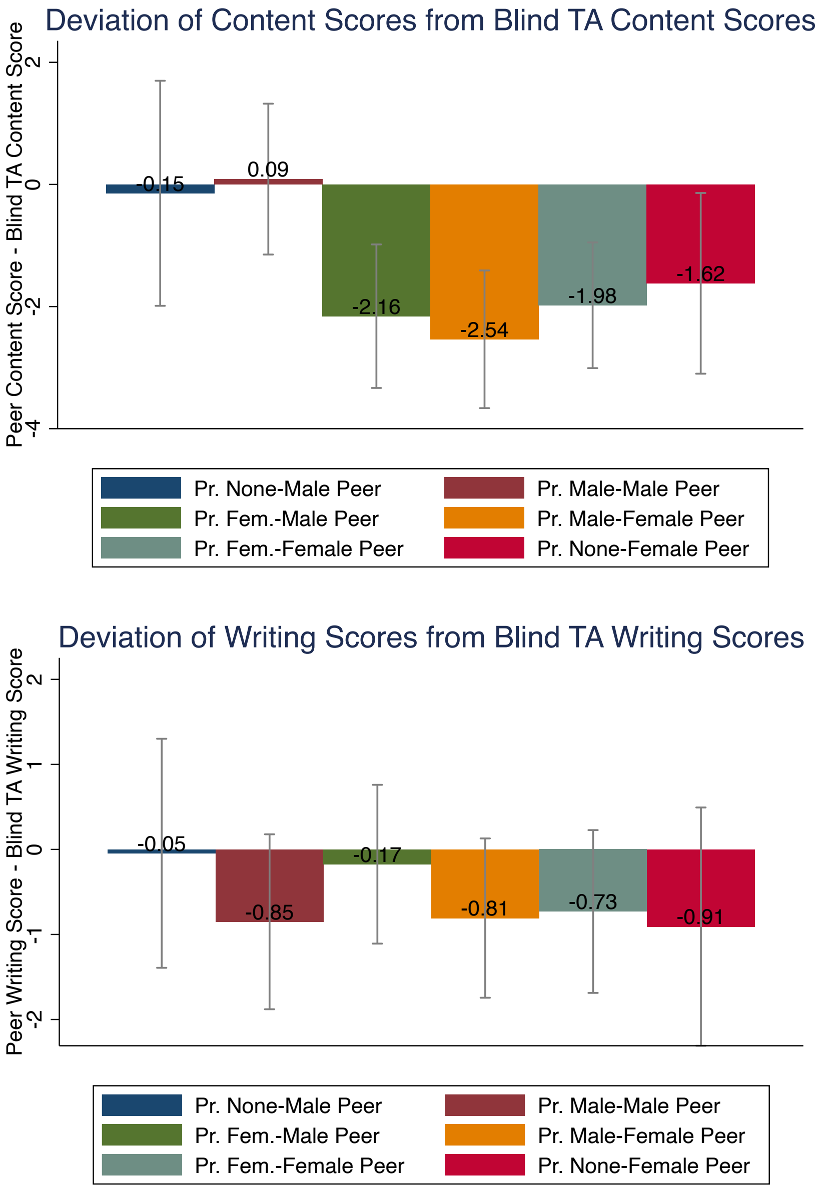
Note: The figures plot the cumulative distribution functions of each subscore, assigned by non-blind peer graders, separately by for male and female peer graders.

Figure 4: Deviations by peer graders' gender and students' predicted gender pairs



Note: These figures show the average deviations of non-blind peer-assigned overall scores from blind TA-assigned overall scores by gender pairs of students and peer graders, where student's gender is predicted by their first name. In the first graph, each student's gender is categorized by a female sounding name if it is predicted as a female name with 90% probability and it is predicted as non-female otherwise. The second graph in this figure further distinguishes the gender pair. Each student's gender is categorized by a female-sounding name, male sounding name, or an ambiguous name. If a name is predicted to be a female or male name with a 90% or higher probability, we define it as female or male sounding name. All other names are defined as none (no prediction).

Figure 5: Deviations in subscores by peer graders' gender and students' predicted gender pairs



Note: These figures show the average deviations of non-blind peer-assigned content and writing subscores from blind TA-assigned scores by gender pairs of students and peer graders, where student's gender is predicted by their first name. Each student's gender is categorized by a female-sounding name, male sounding name, or an ambiguous name. If a name is predicted to be a female or male name with a 90% or higher probability, we define it as female or male sounding name. All other names are defined as none (no prediction).

Table 1: Summary Statistics by Gender

	Female	Male	Gap
	Mean/sd	Mean/sd	b
Panel A: All students enrolled: Selection			
Received a Final Course Grade	0.91	0.92	-0.01
	(0.29)	(0.27)	
Received a Peer Review Score	0.91	0.92	-0.01
	(0.29)	(0.28)	
No, Late, or Incorrect submission	0.23	0.26	-0.03**
	(0.42)	(0.44)	
Not submitted	0.08	0.11	-0.03***
	(0.28)	(0.32)	
Submitted late	0.03	0.02	0.01
	(0.18)	(0.15)	
Peer assigned but no peer grade turned in	0.02	0.01	0.00
	(0.13)	(0.11)	
Homework Submitted in TA bin but not in Peer Bin	0.10	0.11	-0.01
	(0.30)	(0.32)	
Homework Submitted in Peer Bin but not in TA Bin	0.01	0.01	0.00
	(0.10)	(0.09)	
Observations	2148	1752	3900
Panel B: All students enrolled: Performance			
Exam 1	79.41	82.18	-2.78***
	(14.31)	(13.22)	
Exam 2	75.88	77.07	-1.19**
	(15.76)	(13.75)	
Exam 3	77.62	80.94	-3.32***
	(14.85)	(13.68)	
Final Quiz Average	91.88	89.78	2.10***
	(11.15)	(13.61)	
Final Course Grade	81.30	82.07	-0.77*
	(12.43)	(11.69)	
Blind TA Score	88.54	87.03	1.51***
	(15.80)	(16.16)	
Blind TA Content Subscore	86.05	84.21	1.83**
	(20.30)	(20.80)	
Blind TA writing Subscore	96.11	95.70	0.42*
	(6.74)	(6.77)	
Peer Overall Score	87.70	85.58	2.12***
	(15.95)	(17.20)	
Peer Content Sub	85.05	82.60	2.45***
	(20.12)	(21.46)	
Peer Writing Sub	95.72	94.78	0.94**
	(9.87)	(11.53)	
Panel C: Analysis Sample: Performance			
Exam 1	81.39	83.95	-2.56***
	(13.19)	(11.70)	
Exam 2	77.86	78.42	-0.56
	(14.67)	(12.87)	
Exam 3	78.40	81.23	-2.83***
	(14.28)	(13.81)	
Final Quiz Average	92.95	91.39	1.57***
	(9.21)	(10.96)	
Final Course Grade	82.58	83.25	-0.66*
	(11.05)	(10.14)	
Blind TA Score	88.96	87.38	1.58***
	(15.44)	(16.09)	
Blind TA Content Subscore	86.59	84.64	1.95**
	(19.83)	(20.77)	
Blind TA writing Subscore	96.11	95.79	0.33
	(6.80)	(6.67)	
Peer Overall Score	87.78	85.83	1.95***
	(15.93)	(17.01)	
Peer Content Sub	85.15	82.91	2.25***
	(20.09)	(21.21)	
Peer Writing Sub	95.72	94.84	0.88**
	(9.93)	(11.50)	
Observations	1619	1272	2891

Note: Standard deviations are in parentheses. * p<0.10, ** p<0.05, *** p<0.010

Table 2: Balance by Peer Grader's Gender

	Female Peer Mean/sd	Male Peer Mean/sd	Difference b
Female Student	0.56	0.56	0.01
	(0.50)	(0.50)	
Predicted Female Student	0.38	0.38	0.01
	(0.49)	(0.49)	
Predicted Male Student	0.40	0.41	-0.01
	(0.49)	(0.49)	
No Predicted Gender	0.21	0.21	0.00
	(0.41)	(0.41)	
Pr. Of Female Name	0.49	0.48	0.01
	(0.44)	(0.44)	
Female Blind TA	0.43	0.46	-0.03
	(0.50)	(0.50)	
Exam 1	82.39	82.67	-0.28
	(12.62)	(12.62)	
Exam 2	78.02	78.23	-0.21
	(13.92)	(13.90)	
Exam 3	79.65	79.65	-0.00
	(14.37)	(13.84)	
Final Quiz Average	92.24	92.30	-0.06
	(10.16)	(9.90)	
Final Course Grade	82.81	82.96	-0.14
	(10.69)	(10.63)	
Blind TA Score	87.89	88.78	-0.89
	(16.28)	(14.98)	
Blind TA Content Subscore	85.24	86.41	-1.18
	(20.89)	(19.37)	
Blind TA writing Subscore	95.98	95.96	0.02
	(6.77)	(6.72)	
Peer Overall Score	86.09	88.06	-1.97***
	(17.24)	(15.20)	
Peer Content Sub	83.11	85.60	-2.49***
	(21.52)	(19.23)	
Peer Writing Sub	95.18	95.53	-0.36
	(11.42)	(9.51)	
Observations	1670	1221	2891

Note: Standard deviations are in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.010$

Table 3: Balance by Teaching Assistant's Gender

	Female Blind TA Mean/sd	Male Blind TA Mean/sd	Difference b
Female Student	0.55	0.57	-0.02
	(0.50)	(0.50)	
Female Peer Grader	0.56	0.59	-0.02
	(0.50)	(0.49)	
Exam 1	82.37	82.62	-0.25
	(12.70)	(12.56)	
Exam 2	78.40	77.88	0.52
	(13.84)	(13.97)	
Exam 3	79.74	79.58	0.15
	(13.86)	(14.37)	
Final Quiz Average	92.31	92.23	0.08
	(10.11)	(10.01)	
Final Course Grade	82.96	82.81	0.15
	(10.67)	(10.66)	
Peer Overall Score	87.01	86.86	0.15
	(16.66)	(16.26)	
Peer Content Sub	84.27	84.08	0.20
	(20.98)	(20.33)	
Peer Writing Sub	95.38	95.29	0.10
	(10.42)	(10.84)	
Blind TA Score	88.07	88.42	-0.36
	(15.88)	(15.64)	
Blind TA Content Subscore	85.58	85.85	-0.27
	(20.43)	(20.14)	
Blind TA writing Subscore	95.59	96.28	-0.69***
	(8.22)	(5.28)	
Observations	1279	1612	2891

Note: Standard deviations are in parentheses. * p<0.10, ** p<0.05, *** p<0.010

Table 4: Content Sub-Scores: Peer Grader and Student FEs

	PG FEs	PG FEs	Stu. FEs	Stu. FEs
Predicted female*Female Peer Gr.	0.223	0.205	-0.219	-0.145
	(0.831)	(0.829)	(0.939)	(0.942)
Predicted Non-female*Male Peer Gr.	2.063**	2.150**	2.285***	2.286***
	(0.963)	(0.963)	(0.858)	(0.857)
Blind TA Content Subscore	0.786***	0.787***	0.692***	0.693***
	(0.0191)	(0.0192)	(0.0236)	(0.0237)
Blind TA Writing Subscore	0.0883*	0.0871*	0.0528	0.0524
	(0.0451)	(0.0451)	(0.0529)	(0.0532)
Female Blind TA	1.356**	1.379**	0.0167	0.0672
	(0.606)	(0.605)	(0.629)	(0.628)
Graders Content Subscore		-0.0379**		-0.0368**
		(0.0183)		(0.0148)
Graders Writing Subscore		-0.0288		-0.0536
		(0.0518)		(0.0469)
Peer Grader FE	Yes	Yes	No	No
Student FE	No	No	Yes	Yes
Adj. R-squared	0.604	0.604	0.595	0.596
Observations	2793	2793	2793	2793

Note: Standard errors are in parentheses and clustered at student and peer grader level. The dependent variable is peer-assigned content subscores. The estimations in first two columns control for fixed effects for peer graders and the column 3 and 4 control for student fixed effects. * p<0.10, ** p<0.05, *** p<0.010

Table 5: Balance Check: Blind Scores by Gender Pairs

	Blind Score	Blind Content	Blind Writing
Predicted female*Female Peer Gr.	-0.301	-0.371	0.00635
	(0.954)	(1.237)	(0.409)
Predicted Non-female*Male Peer Gr.	0.531	0.808	-0.294
	(0.793)	(1.037)	(0.391)
Student FEs	Yes	Yes	Yes
Observations	2793	2793	2793

Note: Standard errors are in parentheses and clustered at student and peer grader level. The dependent variables are scores assigned by blind expert teaching assistants and varies in each column as indicated by the column names. All estimations control for fixed effects for students. * p<0.10, ** p<0.05, *** p<0.010

Table 6: Writing Sub-Scores: Peer Grader and Student FEs

	PG FEs	PG FEs	Stu. FEs	Stu. FEs
Predicted female*Female Peer Gr.	-0.616	-0.611	-0.526	-0.411
	(0.665)	(0.667)	(0.677)	(0.674)
Predicted Non-female*Male Peer Gr.	-1.511**	-1.470**	-0.348	-0.293
	(0.649)	(0.649)	(0.664)	(0.660)
Blind TA Content Subscore	0.0939***	0.0941***	0.0879***	0.0879***
	(0.0134)	(0.0135)	(0.0155)	(0.0155)
Blind TA Writing Subscore	0.115***	0.113***	0.0849*	0.0823*
	(0.0386)	(0.0386)	(0.0440)	(0.0436)
Female Blind TA	0.411	0.419	0.395	0.410
	(0.468)	(0.469)	(0.519)	(0.520)
Graders Content Subscore		-0.0191		-0.0398***
		(0.0143)		(0.0126)
Graders Writing Subscore		0.0272		0.0592
		(0.0440)		(0.0433)
Peer Grader FE	Yes	Yes	No	No
Student FE	No	No	Yes	Yes
Adj. R-squared	0.153	0.153	0.0241	0.0284
Observations	2793	2793	2793	2793

Note: Standard errors are in parentheses and clustered at student and peer grader level. The dependent variable is peer-assigned writing scores. The estimations in first two columns control for fixed effects for peer graders and the column 3 and 4 control for student fixed effects. * p<0.10, ** p<0.05, *** p<0.010

Table 7: Content Sub-Scores: Peer Grader and Student FEs

	(PG FEs)	(PG FEs)	(PG FEs)	(S FEs)	(S FEs)	(S FEs)
Predicted female*Female Peer Gr.	0.205	2.355*		-0.145	2.140*	
	(0.829)	(1.297)		(0.942)	(1.257)	
Predicted Non-female*Male Peer Gr.	2.150**		2.355*	2.286***		2.140*
	(0.963)		(1.297)	(0.857)		(1.257)
Predicted Female Student		-2.150**				
		(0.963)				
Predicted Non-female Student			-0.205			
			(0.829)			
Female Peer Grader					-2.286***	
					(0.857)	
Male Peer Grader						0.145
						(0.942)
Peer Grader FE	Yes	Yes	Yes	No	No	No
Student FE	No	No	No	Yes	Yes	Yes
All Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2793	2793	2793	2793	2793	2793

Note: Standard errors are in parentheses and clustered at student and peer grader level. The dependent variable is peer-assigned content scores. The estimations in first three columns control for fixed effects for peer graders and the last three column control for student fixed effects. This table aims to show that interaction terms in fixed effects specifications give us the same results directly which can be obtained in two separate specifications. For example, using equation 1, the estimated coefficient of Predicted Non-female*Male Peer grader in the first column (2.150) shows how a same male peer grader grades a student without a female sounding name compared to a student with a female sounding name. This is also obtained in the second column as the estimated coefficient of Predicted Female student (-2.150). Similarly, the estimate for the Predicted female*Female Peer Grader in the first column (0.205) directly shows how a same female peer grader grades a student with a female sounding name compared to a student without a female sounding name. The same result is obtained in third column as the coefficient of Predicted Non-female Student (-0.205). * p<0.10, ** p<0.05, *** p<0.010

Table 8: Content Sub-Scores: Peer Grader and Student FEs

	PG FEs	PG FEs	Stu. FEs	Stu. FEs
Prob of female*Female Peer Gr.	0.465	0.456	-0.521	-0.443
	(0.928)	(0.926)	(0.933)	(0.935)
Prob of Non-female*Male Peer Gr.	1.896*	1.989*	2.423**	2.442**
	(1.040)	(1.038)	(0.999)	(0.999)
Blind TA Content Subscore	0.786***	0.788***	0.692***	0.694***
	(0.0191)	(0.0193)	(0.0236)	(0.0236)
Blind TA Writing Subscore	0.0885*	0.0872*	0.0524	0.0520
	(0.0451)	(0.0452)	(0.0529)	(0.0531)
Female Blind TA	1.377**	1.400**	0.0199	0.0705
	(0.608)	(0.606)	(0.629)	(0.628)
Graders Content Subscore		-0.0377**		-0.0367**
		(0.0183)		(0.0148)
Graders Writing Subscore		-0.0280		-0.0539
		(0.0520)		(0.0469)
Peer Grader FE	Yes	Yes	No	No
Student FE	No	No	Yes	Yes
Adj. R-squared	0.603	0.604	0.594	0.596
Observations	2793	2793	2793	2793

Note: Standard errors are in parentheses and clustered at student and peer grader level. The dependent variable is peer-assigned content subscores. The estimations in first two columns control for fixed effects for peer graders and the column 3 and 4 control for student fixed effects. This table provides a robustness check for the definition of female sounding names. Instead of an indicator variable for a female-sounding name with a certain probability threshold, we use directly the probability of being a female name. * p<0.10, ** p<0.05, *** p<0.010

Table 9: Content Sub-Scores: Peer Grader and Student FEs

	PG FEs	PG FEs	Stu. FEs	Stu. FEs
Predicted female(70)*Female Peer Gr.	0.162	0.145	-0.499	-0.442
	(0.818)	(0.817)	(0.924)	(0.923)
Predicted Non-female(70)*Male Peer Gr.	1.887**	1.938**	2.299***	2.295***
	(0.945)	(0.946)	(0.879)	(0.879)
Blind TA Content Subscore	0.786***	0.787***	0.692***	0.693***
	(0.0191)	(0.0192)	(0.0236)	(0.0237)
Blind TA Writing Subscore	0.0898**	0.0887**	0.0525	0.0521
	(0.0451)	(0.0452)	(0.0528)	(0.0531)
Female Blind TA	1.360**	1.383**	0.0191	0.0696
	(0.607)	(0.606)	(0.629)	(0.629)
Graders Content Subscore		-0.0371**		-0.0365**
		(0.0184)		(0.0148)
Graders Writing Subscore		-0.0293		-0.0540
		(0.0520)		(0.0469)
Peer Grader FE	Yes	Yes	No	No
Student FE	No	No	Yes	Yes
Adj. R-squared	0.604	0.604	0.594	0.596
Observations	2793	2793	2793	2793

Note: Standard errors are in parentheses and clustered at student and peer grader level. The dependent variable is peer-assigned content subscores. The estimations in first two columns control for fixed effects for peer graders and the column 3 and 4 control for student fixed effects. This table provides a robustness check for the definition of female sounding names. We use an indicator variable for a female-sounding name with a 70% probability threshold instead of 90% as in main results. * p<0.10, ** p<0.05, *** p<0.010

Table 10: Writing Sub-Scores: Peer Grader and Student FEs

	PG FEs	PG FEs	Stu. FEs	Stu. FEs
Prob of female*Female Peer Gr.	-0.423	-0.421	-0.470	-0.354
	(0.703)	(0.705)	(0.654)	(0.650)
Prob of Non-female*Male Peer Gr.	-1.660**	-1.610**	-0.474	-0.383
	(0.751)	(0.751)	(0.766)	(0.760)
Blind TA Content Subscore	0.0931***	0.0933***	0.0878***	0.0878***
	(0.0135)	(0.0135)	(0.0155)	(0.0155)
Blind TA Writing Subscore	0.115***	0.113***	0.0851*	0.0825*
	(0.0386)	(0.0385)	(0.0440)	(0.0436)
Female Blind TA	0.410	0.418	0.394	0.409
	(0.467)	(0.468)	(0.519)	(0.520)
Graders Content Subscore		-0.0191		-0.0398***
		(0.0143)		(0.0126)
Graders Writing Subscore		0.0271		0.0592
		(0.0440)		(0.0433)
Peer Grader FE	Yes	Yes	No	No
Student FE	No	No	Yes	Yes
Adj. R-squared	0.152	0.152	0.0241	0.0283
Observations	2793	2793	2793	2793

Note: Standard errors are in parentheses and clustered at student and peer grader level. The dependent variable is peer-assigned writng subscores. The estimations in first two columns control for fixed effects for peer graders and the column 3 and 4 control for student fixed effects. This table provides a robustness check for the definition of female sounding names. Instead of an indicator variable for a female-sounding name with a certain probability threshold, we use directly the probability of being a female name. * p<0.10, ** p<0.05, *** p<0.010

Table 11: Writing Sub-Scores: Peer Grader and Student FEs

	PG FEs	PG FEs	Stu. FEs	Stu. FEs
Predicted female(70)*Female Peer Gr.	-0.336	-0.333	-0.187	-0.0941
	(0.634)	(0.636)	(0.613)	(0.609)
Predicted Non-female(70)*Male Peer Gr.	-1.561**	-1.537**	-0.176	-0.120
	(0.643)	(0.642)	(0.723)	(0.719)
Blind TA Content Subscore	0.0936***	0.0939***	0.0878***	0.0878***
	(0.0134)	(0.0135)	(0.0155)	(0.0155)
Blind TA Writing Subscore	0.113***	0.112***	0.0852*	0.0826*
	(0.0386)	(0.0385)	(0.0440)	(0.0436)
Female Blind TA	0.419	0.427	0.393	0.408
	(0.467)	(0.469)	(0.519)	(0.520)
Graders Content Subscore		-0.0196		-0.0401***
		(0.0143)		(0.0126)
Graders Writing Subscore		0.0279		0.0601
		(0.0441)		(0.0433)
Peer Grader FE	Yes	Yes	No	No
Student FE	No	No	Yes	Yes
Adj. R-squared	0.153	0.153	0.0238	0.0281
Observations	2793	2793	2793	2793

Note: Standard errors are in parentheses and clustered at student and peer grader level. The dependent variable is peer-assigned writing subscores. The estimations in first two columns control for fixed effects for peer graders and the column 3 and 4 control for student fixed effects. This table provides a robustness check for the definition of female sounding names. We use an indicator variable for a female-sounding name with a 70% probability threshold instead of 90% as in main results. * p<0.10, ** p<0.05, *** p<0.010