Journal of Behavioral and Experimental Economics 122 (2026) 102568

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

journal homepage: www.elsevier.com/locate/jbee

# Gender bias in peer performance evaluations☆

Perihan O. Saygin [a,*], Thomas Knight [b]

[a] Department of Applied Economics, Universitat Autònoma de Barcelona, Spain
[b] Department of Economics, Robert F. Lanzillotti Public Policy Research Center, University of Florida, United States of America

ARTICLE INFO

*JEL classification:*
J16
I23

*Keywords:*
Performance evaluation
Peer evaluation
Gender bias

ABSTRACT

Peer performance evaluations are an important determinant of hiring and promotion decisions, but how objective are they? To tackle this question, we measure gender bias in performance evaluations in a large course at a flagship public university. We exploit the random assignments of both peer evaluators and blinded official graders over several essay assignments, where they are incentivized to match official blinded grades. We find that male peer graders assign higher scores to classmates without female-sounding names in content, but lower scores in writing style. Interestingly, we do not find such biases for female graders. Our findings highlight potential challenges of designing fair assessment practices.

## 1. Introduction

Women are underrepresented at 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 McKinsey & 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 blind 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 blind 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 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

☆ 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, Dana Sisak, 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, TOBB Economics, Bank of Spain, Erasmus School of Economics, Universidad Autonoma de Madrid, and Universidad Autonoma de Barcelona, Applied Economics seminars participants. All errors are our own. IRB ID: 202100434 (Study: Discrepancies between Peer Assigned and Instructor Assigned Grades). This study was registered as AEARCTR-0012838. The data used in this paper was collected from the regular implementation of the course syllabus therefore the experiment was registered after the data have been collected.

* Corresponding author.
*E-mail address:* peri.saygin@uab.cat (P.O. Saygin).

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

https://doi.org/10.1016/j.socec.2026.102568
Received 18 June 2025; Received in revised form 29 November 2025; Accepted 6 April 2026
Available online 27 April 2026
2214-8043/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

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 in favor of male-sounding names,[2] and it reverses direction in writing subscores in favor of female-sounding names. Female peers seem to deviate from blind TA-assigned scores more negatively than male peers. However, we do not find any differences in female peers' evaluations based on the predicted gender of the 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. On the other hand, 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. Interestingly, these gender differences against female-sounding names are observed in the content subscores, but the writing subscores do not exhibit similar biases. Instead, male peers seem to favor female sounding names when it comes to writing subscores. This reversal in findings suggests that these gender differences in peer evaluations are consistent with gender stereotypes rather than simply favoring their own group.

Our paper examines an important potential 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.

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 & Rouse, 2000; Milkman et al., 2015; Quadlin, 2018; Sarsons, 2017). Several correspondence experiments document discrimination against women in many contexts (see Bertrand & Duflo, 2017 for an overview). More recently, Kline and Walters (2021) and Kline et al. (2022) find an insignificant difference in contact rates for male and female applicants where they find a significant variation across firms suggesting that some firms favor female applicants while others favor male applicants. Our paper complements these studies. Most importantly, our paper detects some 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 study is distinguished from the existing literature on gender biases in performance evaluations in three important ways. First, we conduct this analysis using data from a real-life classroom setting. Most previous studies have relied on online or lab experiments or audit studies to identify and measure gender biases. Our classroom experiment observes behavior in an actual educational setting, in which both the quality of the written work (by the author) and the quality of the performance evaluation (by the peer reviewer) are incentivized with real payoffs. The subjects of the study are actual students who are maximizing their course grade. Second, in our setting, the author and the peer reviewer have completed the same assignment. This feature allows us to condition our analysis on the peer reviewers' own assignment score. That is, we observe how peer reviewers' own performance affect their evaluations of others' work. Finally, we are able to decompose evaluation scores into different tasks. More specifically, we observe evaluation scores for tasks with objectively correct answers (content subscores) and tasks that are vulnerable to more subjective evaluation (writing subscores).

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., 2020; 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. (2020) 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.

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 gen-

[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.

[3] Some papers have tackled this challenge in other evaluation contexts. For example Card et al. (2020) 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.

[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.

der 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.

Our setting rules 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. Crucially, they are incentivized to provide accurate assessments: their own course grade depends on how closely their peer evaluations align, on average, with the TA's blind evaluations of the same assignment. These incentives are essential—without them, we would be concerned about low effort and inattentive reviews. By creating an environment where all participants are motivated to provide careful and fair assessments, and where other strategic, social, or reciprocal motives are minimized, our design isolates the potential for gender bias more conservatively. Moreover, blind grading is a general policy of the course of which the students are made aware regularly in this course, so that they do not worry about their grades being evaluated by specific TAs (who hold office hours) or the instructor who knows or does not know them. This reduces the chances that peer evaluators would try to infer or "match" any gender-specific stereotypes they imagine the TAs might hold. As a result, the estimates we obtain should be interpreted as a lower bound on gender disparities in peer performance evaluations in more complex real-world settings.

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. (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 (Bordalo et al., 2016, 2019; Coffman, 2014; Sarsons et al., 2021; Stoddard et al., 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 collected data a peer grading exercise 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 data was collected 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.

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

[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 & Bol, 2007; Topping, 1998).

[8] A sample assignment prompt is provided in Fig. A.1.

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

(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).[10] 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. An essential component of our setup is that TA graders neither engage with students[11] nor access any information regarding their records, such as majors or past performance in the course at any point during the semester. Additionally, 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 their first name.[12] 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 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.[13] 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.

The students were not provided information about the identity or gender of the TAs, nor were they provided specific information about the TAs' grading process. For example, students were not told whether a single TA or multiple TAs would grade their work. They were not told if graders were randomly assigned, or if the same grader or graders would evaluate all of their assignments. They were simply told that the instructor and a group of TAs would evaluate their work blindly using the same rubric as they were provided and assign official grades. The official blind grading is communicated to the students regularly in this course, so that they do not worry about their grades being evaluated by specific TAs (who hold office hours) or the instructor who knows or does not know them. Blind grading ensures that neither the TAs nor the instructor can favor or penalize students based on their performance outside of this particular assignment, their attendance and in-class engagement, or other observed student characteristics. This is a general policy of the course of which the students are made aware. This information is included in the syllabus and applies to all assignments, not only to this particular assignment. Because blind grading is a routine practice rather than something introduced for this particular assignment, students do not generally see it as an unusual or strategically informative feature. Instead, they interpret it as the standard overall fairness protection (not particularly for avoiding gender discrimination) used in the course. Given that all students are aware that the official grading is done blindly, peer graders would have no incentive to match an anticipated bias.

It is important to highlight one of the most relevant features of our 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 while ensuring a high participation in the peer evaluation exercise. The purpose of the peer review and the incentive to match TA scores was only in place to encourage students to thoughtfully interact with the correct answers, which were included on the scoring rubric, after the assignment submission deadline. This purpose was communicated clearly to students. Given that peers graders are incentivized to match grades that are given blindly, we consider our results to be conservative. 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 load simultaneously 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

[10] A detailed rubric is provided in the Appendix section in Table A.3.

[11] They do not handle grade disputes or students' questions about their grades and they also do not hold any office hours.

[12] Peer graders could possibly look at the picture of the student, if available. However, most students chose not to upload their pictures into Canvas. Only around 10 percent of students had pictures, and even when they did, the images were tiny, approximately the size of a dime, and lacked an option to enlarge them. Furthermore, these images often failed to convey informative details about the students' identities, either because they included multiple individuals or none at all. Moreover, we do not find significant differences in the availability of pictures by gender.

[13] The threshold was +/−5 points to receive credit.

**Table 1**
Summary statistics by predicted gender

| | Predicted female<br>Mean/sd | Predicted non-female<br>Mean/sd | Gap<br>b |
|---|---|---|---|
| Analysis Sample: Gender Predictions and Performance | | | |
| Female Student | 1.00 | 0.29 | 0.71*** |
| | (0.00) | (0.45) | |
| Prob. of Female Name | 0.97 | 0.18 | 0.79*** |
| | (0.02) | (0.28) | |
| Exam 1 | 81.23 | 83.26 | −2.03*** |
| | (13.61) | (11.87) | |
| Exam 2 | 77.36 | 78.43 | −1.07* |
| | (15.07) | (13.06) | |
| Exam 3 | 77.86 | 80.66 | −2.80*** |
| | (15.10) | (13.42) | |
| Final Quiz Average | 93.28 | 91.65 | 1.63*** |
| | (8.03) | (10.92) | |
| Final Course Grade | 82.37 | 83.14 | −0.77* |
| | (11.26) | (10.07) | |
| Blind TA Score | 89.50 | 87.52 | 1.98*** |
| | (14.90) | (16.14) | |
| Blind TA Content Subscore | 87.26 | 84.79 | 2.47*** |
| | (19.04) | (20.90) | |
| Blind TA writing Subscore | 96.28 | 95.85 | 0.43* |
| | (6.59) | (6.70) | |
| Peer Overall Score | 87.82 | 86.30 | 1.52** |
| | (15.74) | (16.82) | |
| Peer Content Sub | 85.19 | 83.45 | 1.74** |
| | (19.89) | (21.04) | |
| Peer Writing Sub | 95.75 | 95.00 | 0.76* |
| | (10.14) | (11.02) | |
| Observations | 1063 | 1730 | 2793 |

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

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. Table 1 shows the summary statistics for the full sample (Panel A and B) and the sample of the analysis (Panel C). While, as expected, we observe a slight positive selection into the sample of the analysis compared to the full sample, we do not see a difference in terms of the observed differences among students with female and male sounding names.

Our experimental design allows us to assess gender bias by using inferred gender from students' first names. As described in the previous section, we use the predicted gender in our analysis rather than the gender reported in administrative data.[14] Using genderize.io, we predict the gender probabilities for all first names. In our descriptive statistics, we define the name as female-sounding if the gender is predicted as female with greater than 90% probability. Notice that this is an indicator variable flagging the names that are highly likely to be perceived as female. This definition may misclassify a significant share of female students with relatively more gender-neutral names. In order to test for the robustness of our results to different definitions of female-sounding names, we also used a 70% threshold. However, this definition still needs to rely on a classification based on a threshold which does not rule out the concern for misclassification. Therefore, later in the analysis, we show results relying on the probability of a name being associated with a female taking a value between 0 and 1, rather than an indicator variable.[15]

In order to summarize our data, considering the differences across students with and without female-sounding names, we provide the summary statistics for the full sample and the sample of the analysis in Table A.1. In panel A, we report the descriptive statistics by predicted gender for 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 students with and without female sounding names. On the other hand, students without female-sounding names 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 A.1. We report the same statistics for the restricted sample used in our analysis in panel C. Students with female-sounding names perform better than students without female-sounding names on all assessments, except for the three high-stakes exams.[16] There are small differences between the restricted and full

[14] It is reasonable to assume that peers do not know the author, because our experimental setting is a large and mostly online (except for exams) introductory course. Thus, in the overwhelming majority of cases, the peer evaluator's rely solely on an author's name to identify their gender.

[15] Results are similar regardless of the threshold or use of an indicator or a probability measure.

sample. The restricted sample is slightly more positively selected in terms of performance on assessments, but the gender differences in TA-assigned scores and peer scores and the gender composition appear similar.

Table 1 reveals that female students' names are predicted with less certainty in our sample. Female students in our sample are more likely to have gender-neutral names. All students who are predicted to be female from their first names are also reported as female in the administrative data. However, 29% of students whose names are not identified as female-sounding with 90% or greater probability are reported as female in the administrative data. For example, even in our conservative restriction of 90% probability, we predict 5.34% of female students as male-sounding names.[17]

In Table A.4, we report our summary statistics by students' gender as reported in administrative data. The performance differences by gender are quite similar to the differences by the gender that is predicted from their first names. This table provides consistent observations regarding the accuracy of predicting gender from first names. Female students are more likely to be classified as having an unknown gender. 68% of female students are predicted to be female, 4% are predicted to be male, and the remaining 28% are classified as unknown. 87% of male students are predicted to be male, and no male student was predicted to be female. The remaining 13% of male students have first names with no gender prediction.

In our analysis, we focus on female-sounding names, precisely because none of those names are associated with a single male student. Given that most of the students are unlikely to know each other in a large, online course, the effect of having a female-sounding first name is suggestive of gender bias against women. We run multiple robustness analyses. We relax the strictness of the gender prediction requirement by reducing the probability threshold to 70% classifying a name as female-sounding. Even after relaxing this threshold, we find no male students in our sample have a female-sounding name. We also conduct the analysis using the probability of being a female name rather than constructing a binary female-sounding name indicator. Our results remain consistent across different measures as we present in the next section and the appendix.

One of the advantages of our setup is that the peer evaluators are randomly assigned to grade the assignments. In order to provide evidence for this randomization, 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 or name associated 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 2**
Balance by peer grader's gender.

| | Female peer Mean/sd | Male peer Mean/sd | Difference b |
|---|---|---|---|
| Female Student | 0.56 | 0.56 | 0.00 |
| | (0.50) | (0.50) | |
| Predicted Female Student | 0.38 | 0.38 | 0.00 |
| | (0.49) | (0.49) | |
| Predicted Male Student | 0.41 | 0.41 | −0.00 |
| | (0.49) | (0.49) | |
| No Predicted Gender | 0.21 | 0.21 | −0.00 |
| | (0.41) | (0.41) | |
| Predicted Female Student 70 | 0.45 | 0.44 | 0.01 |
| | (0.50) | (0.50) | |
| Predicted Male Student 70 | 0.47 | 0.47 | −0.00 |
| | (0.50) | (0.50) | |
| No Gender Predicted 70 | 0.08 | 0.09 | −0.01 |
| | (0.28) | (0.29) | |
| Prob. of Female Name | 0.49 | 0.48 | 0.01 |
| | (0.44) | (0.44) | |
| Female Blind TA | 0.43 | 0.46 | −0.02 |
| | (0.50) | (0.50) | |
| Exam 1 | 82.45 | 82.53 | −0.08 |
| | (12.56) | (12.66) | |
| Exam 2 | 77.95 | 78.12 | −0.17 |
| | (13.87) | (13.87) | |
| Exam 3 | 79.68 | 79.48 | 0.20 |
| | (14.35) | (13.86) | |
| Final Quiz Average | 92.26 | 92.29 | −0.03 |
| | (10.05) | (9.80) | |
| Final Course Grade | 82.82 | 82.88 | −0.06 |
| | (10.57) | (10.51) | |
| Blind TA Score | 87.91 | 88.77 | −0.86 |
| | (16.21) | (14.99) | |
| Blind TA Content Subscore | 85.25 | 86.38 | −1.13 |
| | (20.82) | (19.43) | |
| Blind TA writing Subscore | 96.01 | 96.02 | −0.01 |
| | (6.75) | (6.54) | |
| Peer Overall Score | 86.08 | 87.96 | −1.88*** |
| | (17.19) | (15.28) | |
| Peer Content Sub | 83.10 | 85.48 | −2.38*** |
| | (21.47) | (19.35) | |
| Peer Writing Sub | 95.13 | 95.49 | −0.36 |
| | (11.46) | (9.58) | |
| Observations | 1606 | 1187 | 2793 |

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

Table A.2 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.[18] 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 Fig. 1. 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

[16] This observation is consistent with previous findings on gender differences in performance under pressure and when stakes are high.

[17] These names include Alexis, Barlow, Cameron, Jesse, Jordan, Lotanna, Michel, Raleigh, Reagan, Reece, Ruihan, Rylan,Tyler, Willmar, Zekai. Note that, no-female sounding names include male-sounding names and names with no gender prediction. The majority of these names are gender-neutral names that are popular in the birth cohort of the students in our sample such as Taylor, Riley, Bailey, Avery, Sam, Alec, Alex, Miles, Madison, Morgan, Ashley, Ellis, Collin, Tori, Skyler, Jude, and Devon.

[18] In Fig. A.2 we show that the average scores assigned by TAs by gender pairs of students and TAs, where student's gender comes from the administrative data to make sure that male or female TAs grade female or male students differently.

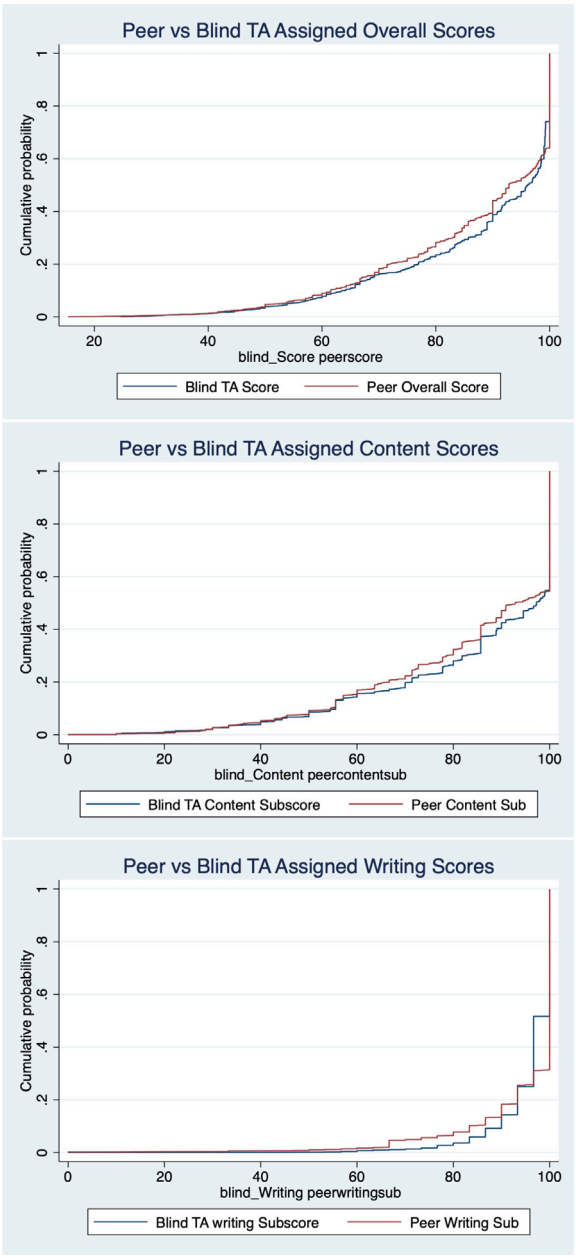


**Fig. 1.** CDF of TA-assigned and peer-assigned grades.
Note: The figures present the cumulative distribution functions for overall scores, content subscores, and writing subscores, separately for blind TA-assigned scores and non-blind peer scores.

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 Fig. 2. 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. Fig. 3 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.[19] 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. Fig. 4 shows the deviations between TA-assigned and peer scores, by gender pairs, for the content and writing subscores. It suggests that the differences are mostly 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 except for the assignments of male-sounding names graded by male peers.

These findings suggest that peer evaluations often deviate negatively from the official grades. Female peers deviate even more negatively, compared to male peers. Female peers do not exhibit any significant differences with respect to the author's predicted gender. However, male peers deviate negatively in content score only for authors with female-sounding names. They deviate negatively in writing scores only for authors with male-sounding names. In the next section, we analyze these differences in more detail.

## 4. Methodology and results

We study how peers evaluate each others' work when they are incentivized to make a careful evaluation. More specifically, we ask whether female-sounding names are evaluated any differently and if the gender of the peer evaluator seem to affect these evaluations. In this section, we lay out an empirical approach for answering these research questions.

First, we rely on a specification that predicts a peer evaluator's assessment of a student from the candidate's predicted gender controlling for the scores assigned by the blind TAs. We cluster standard errors at the student level. Formally, we predict a score assigned by a peer $j$ to a student $i$ on assignment $t$:

$$\text{Peer Score}_{ijt} = \beta_1 \text{Stu F}_{ij} + \beta_2 \text{Peer F}_{ij} + \beta_3 X'_{it} + \beta_4 Z_{jt} + \mu_j + \epsilon_{ijt} \quad (1)$$

where $\text{Stu F}_{ij}$, is a dummy variable that takes the value 1 if the student $i$ has a female-sounding name with at least 90% probability. $\text{Peer F}_{ij}$ is a dummy variable that takes the value 1 if the peer score is given by a female peer grader. $X'_{it}$ includes the TA-assigned score that student $i$ receives on assignment $t$ to serve as a benchmark for the performance of the student $i$. $Z'_{jt}$ includes peer grader $j$'s own TA-assigned scores on assignment $t$, and $\mu_j$ represents the peer grader fixed effects.

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 and partially correct answers. Next, we turn to writing subscores where the rubric provides less clear guidance. In Table 3 we report the results from different specifications of Eq. (1) for both the content and writing subscores. In the first column, our specification controls only for the TA-assigned scores. In the second column, we add an indicator variable for female peer evaluators. The third column presents the results with peer grader fixed effects. Lastly, we also add peer evaluators' own scores assigned by the TAs to see if their own performance affects peer evaluators' assessments. We repeat the same specifications for writing scores in the remaining columns, in the same order. We also present results from predicting the deviation of peer-assigned scores from TA-assigned blind scores as the dependent variable. These results are presented in the Appendix Table A.6.[20]

[19] 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.

[20] In these regressions, we do not control for TA-assigned blind scores. If we did control for blind scores, it would yield to the same exact estimates as in Table 3. We keep the predictions of peer-assigned scores conditional on

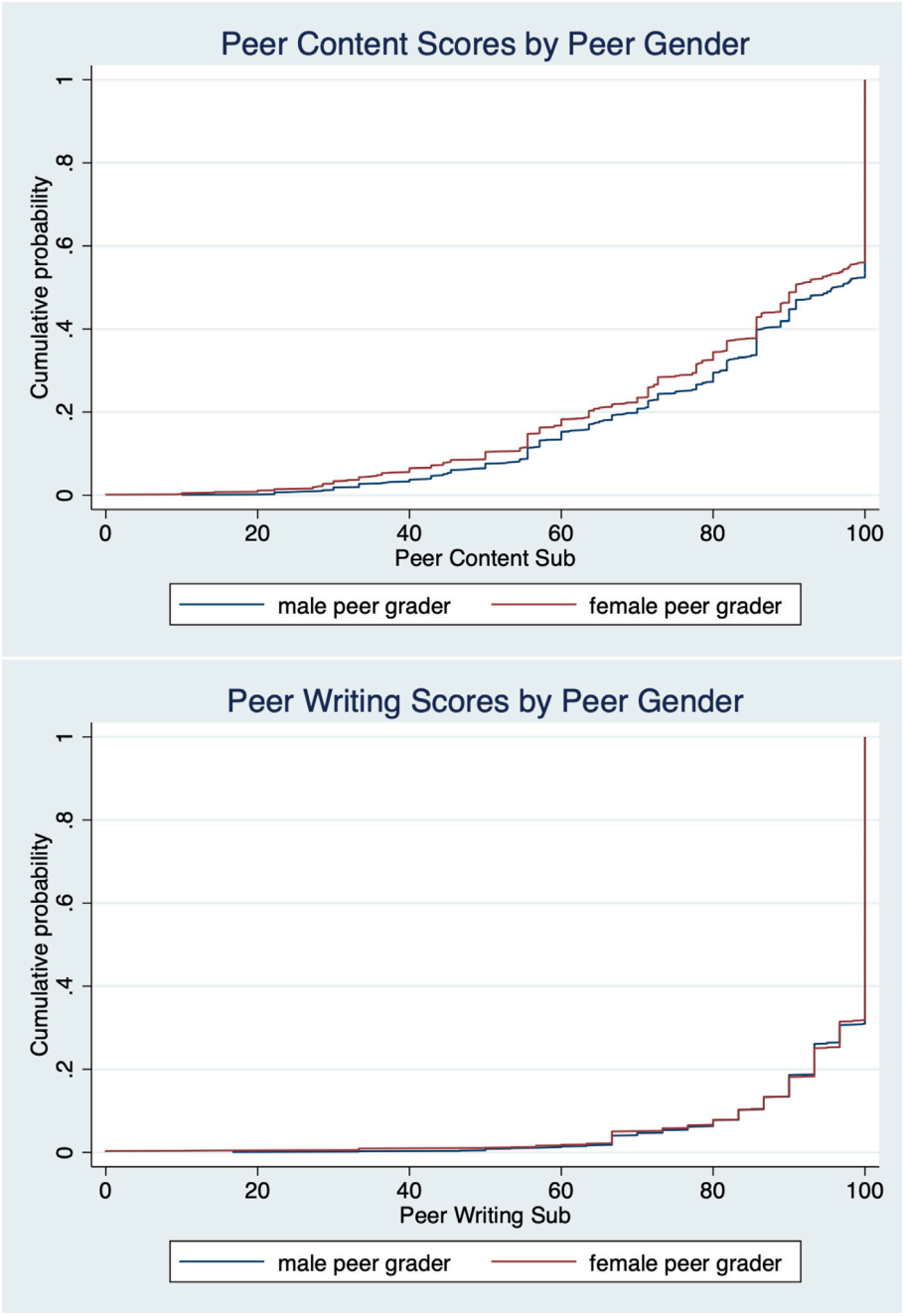


**Fig. 2.** CDF of subscores assigned by peers by peer grader's gender.
Note: The figures plot the cumulative distribution functions of each subscore, assigned by non-blind peer graders, separately by for male and female peer graders.

Across all specifications, we find a small negative (positive) effect for female-sounding names in content (writing) scores, but these estimates are statistically insignificant. That is, controlling for the blindly assigned scores, female-sounding names do not receive significantly better or worse evaluations from their peers in content or writing scores.

As a robustness check for our predicted gender measure, we run the same analysis with alternative definitions of predicted gender. These results rely on the female sounding names with at least 90% probability. This is an indicator variable marking names that are highly likely to be perceived as female. We tested whether our results remain similar with a lower threshold, namely 70%, for the same female-sounding name indicator. These results are reported in the Appendix in Tables A.5. Both of these measures are indicator variables potentially misclassifying gender-neutral names that are associated with female students. In order to avoid this measurement concern, we also repeated the analysis by using the probability of each name being associated with a female in Table 4. Using the probability taking value between 0 and 1, avoids this potential concern and predicts the effect of increased probability of the name being associated with a female student. When we repeat these estimations with alternative definitions of predicted gender, our results remain mostly similar, but we obtain marginally

blind scores as our main specification primarily because deviations may relate to actual performance. Higher performing students will be always more likely to have negative deviations. Indeed, these deviations tend to be larger for high-performing students as shown on Fig. 1.

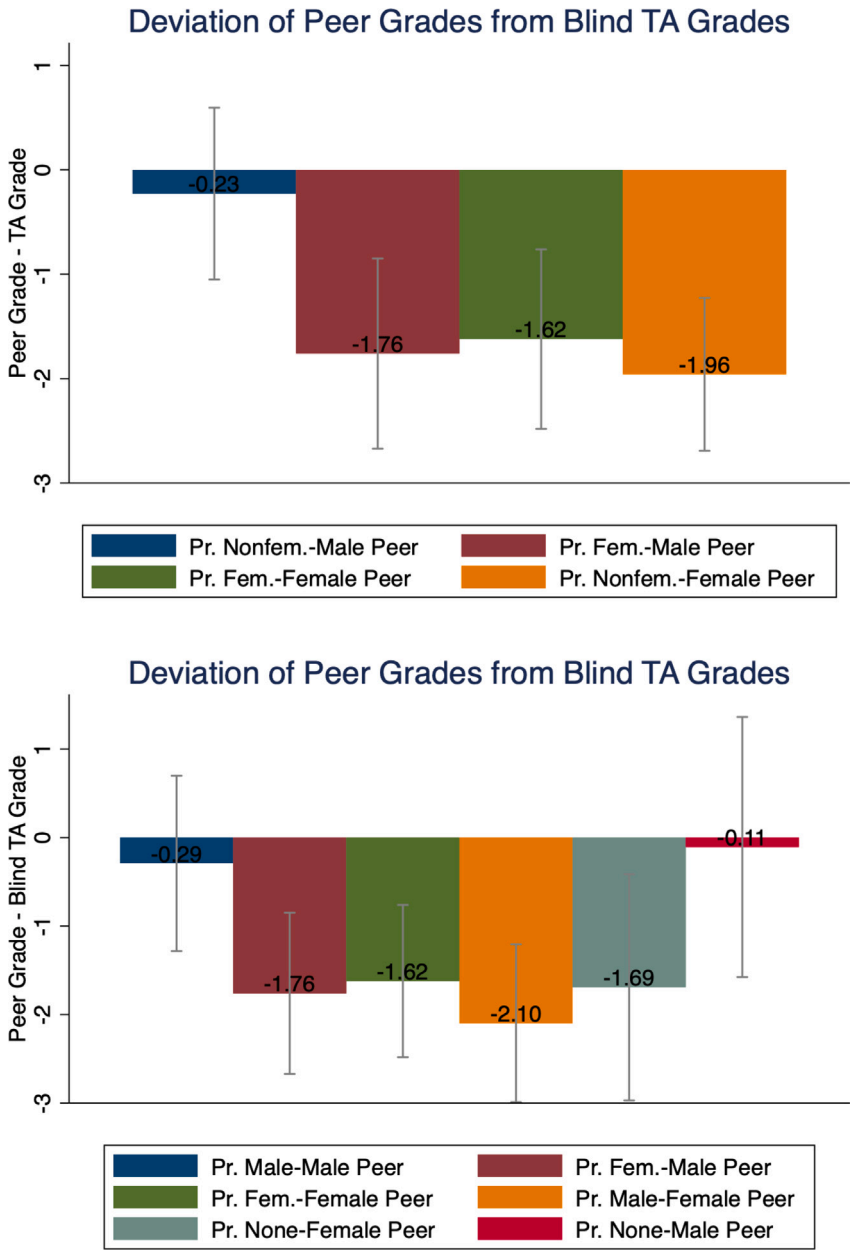


**Fig. 3.** Deviations by peer graders' gender and students' predicted gender pairs.
Note: These figures show the average deviations of non-blind peer-assigned overall scores from blind TA-assigned overall scores by gender pairs of students and peer graders, where student's gender is predicted by their 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).

significant positive point estimates in the writing scores. This suggests that, if anything, female-sounding names are evaluated more favorably when it comes to the writing subscores.

When we turn our attention to the comparison of scores assigned by female and male peers, we observe a significant negative estimate when an assignment is evaluated by a female peer. This is consistent with the observations in Figs. 3 and 4 that we discussed in the previous sections. Peer evaluators tend to deviate negatively from the TA-assigned scores, and the deviation of female peer assigned scores is even more negative, particularly in the content subscores. The gender difference in this deviation does not seem to change once we control for the peer evaluator's own performance on the same assignment. Based on these results, female peers are not tougher evaluators *per se*, as we find this difference only in the content subscores.[21]

Our first set of findings suggests that there is no significant difference in scores assigned to authors with female-sounding names, compared to authors without female-sounding names overall. Next, we examine whether female and/or male peers evaluate authors with female-sounding names any differently in content or writing subscores.

[21] While our set up is not suitable to explore the channels that explain the reasons for female peers being even more critical, it is possible that women have more negative expectations in content evaluations.

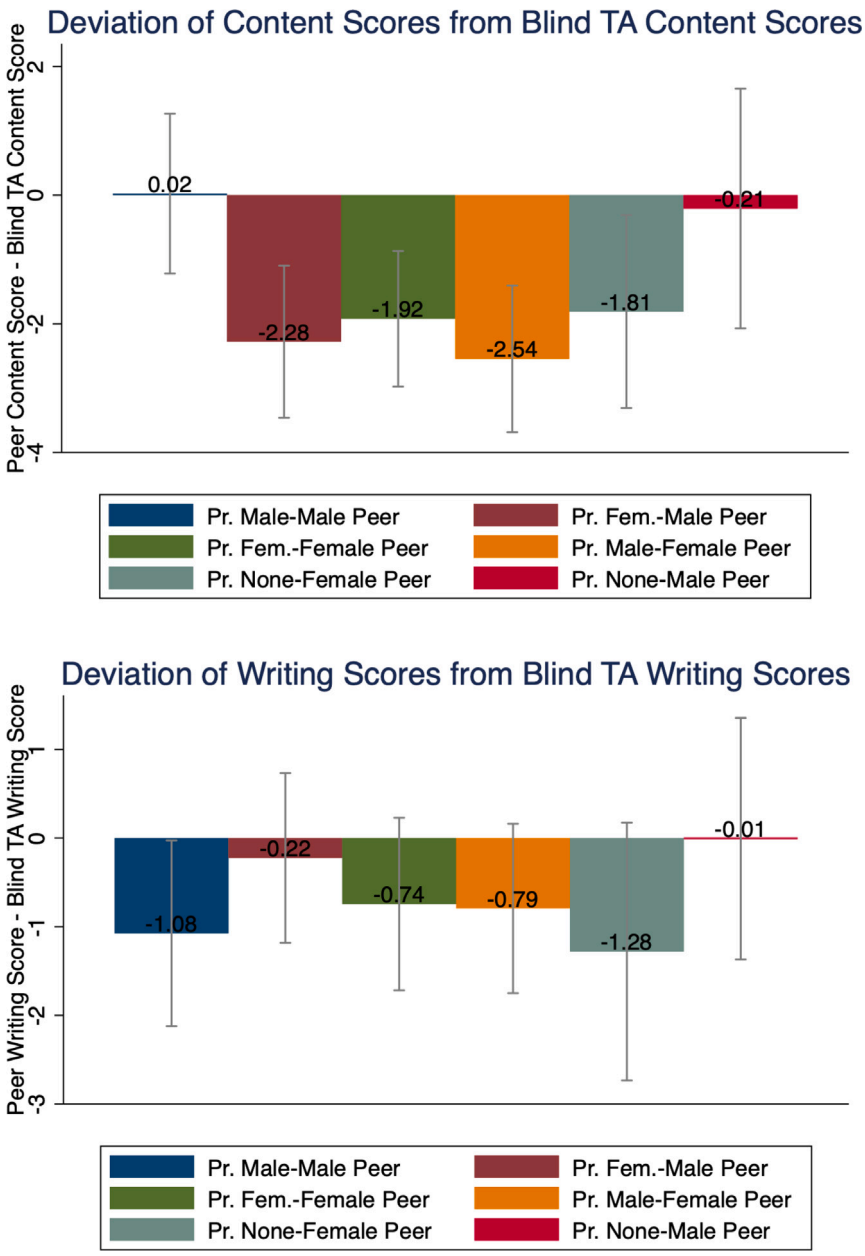


**Fig. 4.** Deviations in subscores by peer graders' gender and students' predicted gender pairs.
Note: These figures show the average deviations of non-blind peer-assigned content and writing subscores from blind TA-assigned scores by gender pairs of students and peer graders, where student's gender 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).

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 versus a randomly assigned student without a female-sounding name.

$$\text{Peer Score}_{ijt} = \beta_1 \text{Stu Peer M}_{ij} + \beta_2 \text{Stu Peer F}_{ij} + \beta_3 X'_{it} + \beta_4 Z_{jt} + \mu_j + \epsilon_{ijt} \quad (2)$$

where $\text{Peer Score}_{ijt}$ 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 $\text{Stu Peer M}_{ij}$, 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 $\text{Stu Peer F}_{ij}$, 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 $\mu_j$ 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:

$$\text{Peer Score}_{ijt} = \beta_1 \text{Stu Peer M}_{ij} + \beta_2 \text{Stu Peer F}_{ij} + \beta_3 X'_{it} + \beta_4 Z_{jt} + \nu_i + \epsilon_{ijt} \quad (3)$$

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

**Table 3**
Content and writing sub-scores: Are female-sounding names evaluated more/less favorably?

| | Content | Content | Content | Content | Writing | Writing | Writing | Writing |
|---|---|---|---|---|---|---|---|---|
| Predicted Female Student | −0.213 | −0.204 | −0.821 | −0.243 | 0.479 | 0.481 | 0.280 | 0.474 |
| | (0.492) | (0.492) | (0.619) | (0.491) | (0.380) | (0.380) | (0.472) | (0.381) |
| Female Peer Grader | | −1.510*** | | −1.453*** | | −0.257 | | −0.227 |
| | | (0.529) | | (0.528) | | (0.380) | | (0.379) |
| Blind TA Content Subscore | 0.774*** | 0.773*** | 0.786*** | 0.774*** | 0.0951*** | 0.0949*** | 0.0942*** | 0.0950*** |
| | (0.0176) | (0.0176) | (0.0190) | (0.0176) | (0.0124) | (0.0124) | (0.0134) | (0.0124) |
| Blind TA writing Subscore | 0.0901** | 0.0907** | 0.0838* | 0.0888** | 0.0978*** | 0.0979*** | 0.113*** | 0.0953*** |
| | (0.0390) | (0.0389) | (0.0452) | (0.0389) | (0.0368) | (0.0368) | (0.0384) | (0.0366) |
| Graders Content Subscore | | | | −0.0408*** | | | | −0.0296*** |
| | | | | (0.0127) | | | | (0.0103) |
| Graders Writing Subscore | | | | −0.0162 | | | | 0.0450 |
| | | | | (0.0349) | | | | (0.0327) |
| Peer Grader FE | No | No | Yes | No | No | No | Yes | No |
| Observations | 2793 | 2793 | 2793 | 2793 | 2793 | 2793 | 2793 | 2793 |
| R-squared | 0.587 | 0.588 | 0.731 | 0.590 | 0.0420 | 0.0422 | 0.426 | 0.0451 |

Note: Standard errors are in parentheses and clustered at student level. The dependent variable is peer-assigned content subscores in the first four columns and writing subscores in the last four columns. The estimations in the 3rd and 7th columns control for fixed effects for peer graders. * p < 0.10, ** p < 0.05, *** p < 0.010.

**Table 4**
Content and writing sub-scores: Are female-sounding names evaluated more/less favorably? - Using probability of female names.

| | Content | Content | Content | Content | Writing | Writing | Writing | Writing |
|---|---|---|---|---|---|---|---|---|
| Prob Of Female Name | 0.216 | 0.231 | −0.580 | 0.202 | 0.743* | 0.745* | 0.444 | 0.737* |
| | (0.550) | (0.550) | (0.682) | (0.549) | (0.427) | (0.426) | (0.521) | (0.427) |
| Female Peer Grader | | −1.513*** | | −1.456*** | | −0.260 | | −0.230 |
| | | (0.530) | | (0.528) | | (0.379) | | (0.379) |
| Blind TA Content Subscore | 0.773*** | 0.772*** | 0.786*** | 0.773*** | 0.0947*** | 0.0945*** | 0.0939*** | 0.0945*** |
| | (0.0177) | (0.0177) | (0.0191) | (0.0176) | (0.0125) | (0.0124) | (0.0134) | (0.0124) |
| Blind TA writing Subscore | 0.0896** | 0.0902** | 0.0840* | 0.0883** | 0.0978*** | 0.0979*** | 0.113*** | 0.0954*** |
| | (0.0390) | (0.0389) | (0.0452) | (0.0389) | (0.0368) | (0.0368) | (0.0384) | (0.0366) |
| Graders Content Subscore | | | | −0.0407*** | | | | −0.0296*** |
| | | | | (0.0128) | | | | (0.0103) |
| Graders Writing Subscore | | | | −0.0154 | | | | 0.0447 |
| | | | | (0.0349) | | | | (0.0327) |
| Peer Grader FE | No | No | Yes | No | No | No | Yes | No |
| Observations | 2793 | 2793 | 2793 | 2793 | 2793 | 2793 | 2793 | 2793 |
| R-squared | 0.587 | 0.588 | 0.731 | 0.590 | 0.0425 | 0.0427 | 0.426 | 0.0455 |

Note: This table replicates the Table 3 by using the probability of the first name being a female name instead of an indicator for a female-sounding name with 90 percent probability. Standard errors are in parentheses and clustered at student level. The dependent variable is peer-assigned content subscores in the first four columns and writing subscores in the last four columns. The estimations in the 3rd and 7th columns control for fixed effects for peer graders. * p < 0.10, ** p < 0.05, *** p < 0.010.

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.

Table 5 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 Eq. (2) and Eq. (3) give the same results as if each of them were estimated as two separate specifications: one in which Stu Peer $F_ij$ is kept and Stu Peer $M_ij$ is replaced by a student's predicted gender dummy variable, and another in which Stu Peer $M_ij$ is kept and the Stu Peer $F_ij$ 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 Appendix in Table A.7. Using Eq. (2), the result on Stu Peer $M_ij$ 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 5, 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

**Table 5**
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 6**
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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$.

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 6 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 7 presents the results for the writing subscores using the same specifications and format as Table 5. 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 opposite 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 7 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.

An alternative explanation for these findings is that disparities in grading proficiency between TAs and peers might manifest as gender bias. Even though blinded TAs are not given access to student characteristics or names, they may still infer certain information from the SWAs during grading. For instance, TAs might attribute different importance to various skills or writing styles (or exhibit better proficiency in assessing certain skills due to expertise) compared to peer graders. If these skills or characteristics of SWAs correlate with students' genders, our comparisons between TA-assigned scores and peer-assigned scores could appear as gender disparities. In our case, disparities would be driven by male peer evaluators who are less inclined to recognize skills or positively evaluate essay characteristics that are more commonly associated with women. However, we conjecture that it is unlikely that this is primary explanation for our findings, because the bias reverses in writing scores in favor of women. The alternative hypothesis would require that TAs are more inclined to recognize certain skills more frequently associated with women in content scores but less so in writing skills.

Our findings suggest that male graders may favor male authors based on their gender. However, an alternative interpretation could be that this preference stems from alignment with the content of their

**Table 7**
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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$.

responses. Male evaluators might award higher grades to male authors, because they appreciate their content and writing style, finding it more similar to their own. We cannot reject either of these explanations as the primary driver of our results. The data does not allow us to disentangle these two motivations for biased performance evaluations. Either way, regardless of whether male peer evaluators are less able to recognize quality in the content of female-authored assignments or are simply more likely to appreciate the content of other male-authored essays, they are systematically less generous in their evaluations to female students and it appears that such an affinity is not mirrored in the scores of female students evaluated by female peers.

One potential concern is that our results may be driven by the incentive for peer graders to match official grades that they expect are already biased. If students expect biases in TA-assigned grades, they could try to replicate those biases to earn a higher grade themselves. As mentioned earlier in the experimental design section, however, peer graders were told that the official grading was done blindly. We do not expect our results to be driven by an intention to match an expected bias among the TAs. Moreover, if this was the main driver of our results, it would imply that only male peer graders are trying to match such expected biases and not female peer graders.

Tables 5 and 7 also provide insights into the general validity of peer grading. We find that content subscores demonstrate considerably more validity than writing subscores. In Table 5, 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 7, 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. These results are similar to what we observe in Table 3 where we also report results without any fixed effects.

Finally, in Tables 5 and 7, 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 the Appendix Table A.8 and Table A.9.

As we explained earlier, we also tested for potential measurement concern due to a missclassification by the indicator variables relying on a threshold. Instead of using these indicator variables to identify a female-sounding name, we used the probability of being female as an alternative measure in Tables 8 and 9 where we do not find significantly different results.

We do not randomly assign female-sounding names to homework submissions; rather, we use students' actual names. For this reason, one might question whether our results could be driven by female-sounding names that signal other characteristics. For example, one might ask if female-sounding names are disproportionately more likely to be associated with a particular ethnicity or race. While we cannot offer a formal analysis to rule out this possibility, it seems highly improbable. We find that the differential evaluation of authors with female-sounding names by male peers reverses when it comes to the writing scores. We would expect this differential evaluation of female-sounding names to remain the same (or at least not reverse) if female-sounding names were disproportionately associated with a particular ethnicity or race. Additionally, the most common female-sounding names in our sample do not appear non-White or international at first glance.[22]

Because peer evaluators observe students' actual names, as opposed to randomly assigned experimental names, we cannot disentangle whether the identified effects are attributable to the observed names or gender-based differences in the assignment submissions. For

[22] They are Abigail, Amanda, Anna, Camila, Chloe, Claudia, Daniella, Elizabeth, Emily, Gabriella, Grace, Hanna, Isabella, Jessica, Julia, Leah, Maria, Megan, Olivia, Paige, Rachel, Samantha, Sara(h), Sofia, and Victoria.

**Table 8**
Content sub-scores: Peer grader and student FEs - Using probability of female names.

| | 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**
Writing sub-scores: Peer grader and student FEs - Using probability of female names.

| | 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 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. 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$.

example, when male peer evaluators favor assignments written by male students, is this effect attributable to unmeasured similarities in how male students compose their submissions and what they expect when conducting a peer evaluation? Because we only observe this effect in the content subscores, and the effect reverses direction in the writing subscores, it is more likely that the effect is driven by the gender-related stereotypes that are triggered by students' names. Additionally, we do not observe such in-group favoring among female peer evaluators.

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.[23] 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

[23] 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.

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.

Because female peer graders do not exhibit similar behavior, female authors, on average, receive harsher evaluations from their male peers when performing male-stereotypical tasks. These findings are particularly compelling, because they were obtained in an experimental 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 against female-sounding names in content subscores among male evaluators. 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.

Interestingly, these differences reverse when it comes to writing subscores. Male peer graders assign approximately 1.5-point lower scores in the writing subscore to students without female-sounding names, compared to students with female-sounding names. Observing a clear gender bias against female-sounding names in the content subscores and a smaller but reversed gender bias in favor of female-sounding names in the writing subscores is suggestive evidence that the bias stems from gender stereotypes, 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 stereotypes.

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.

Two aspects of our design merit careful discussion: (i) the incentives that tie peer evaluators' grades to the accuracy of their assessments, and (ii) the informational asymmetry created by blind TA grading combined with peer evaluators' access to the author's first name. The incentive structure is crucial for ensuring that peer reviewers participate, exert effort, and provide thoughtful, careful evaluations. Absent these incentives, low-effort grading could obscure systematic differences by gender as well as low levels of participation. By motivating peer evaluators to align their assessments with those of the blind TAs, the design helps us isolate gender differences that persist even when reviewers are incentivized to be accurate and objective.

The informational asymmetry also has important implications. Because students know that TA graders are blind, they have no reason to believe that TA evaluations rely on gender-based stereotypes, and thus no strategic incentive to match such biases. At the same time, peer evaluators do observe the author's first name, which allows us to capture any gender-based differences that arise when this information is available. This combination of blind official grading and non-blind peer grading therefore helps us identify gender disparities in a setting where reviewers cannot rationally expect stereotypes to help them match the TA's assessments.

These features imply that our estimates should be interpreted as conservative. In many real-world peer evaluation settings—such as workplace performance reviews, academic refereeing, or committee-based assessments—evaluators may believe that identity-related information is available to and potentially used by official decision-makers. In such environments, peer evaluators might strategically match or even amplify perceived biases of those evaluators, or rely more heavily on stereotypes when uncertainty is high. In contrast, in our setting, students know that TA graders are blind and thus have no reason to believe that TA assessments depend on gender. As a result, any gender differences we detect reflect peer evaluators' own stereotypes or beliefs, rather than attempts to anticipate or replicate perceived TA biases. If TA graders in our context were not blind, or if peer evaluators believed that TAs observed student identity, then strategic alignment with perceived TA stereotypes could plausibly intensify these disparities. For this reason, the biases documented in our experiment represent a lower bound: they are the gender disparities that persist even when (i) official evaluators are blind, (ii) matching TA stereotypes

is neither feasible nor rewarded, and (iii) peer reviewers are motivated to provide accurate assessments. In this sense, our controlled environment likely understates the magnitude of gender bias that may arise in more complex real-world peer evaluation settings where additional strategic or social motives are present.

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. We make these observations in an experimental setting. Isolating and measuring gender biases in a real-world workplace is impractical or entirely infeasible. While concerns surrounding external validity always arise with such experiments, our experimental setting provides an objective benchmark for comparison, which is rarely available in the actual workplace.

## CRediT authorship contribution statement

**Perihan O. Saygin:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Thomas Knight:** Writing – review & editing, Resources, Project administration, Data curation, Conceptualization.

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.socec.2026.102568.

## Data availability

The data used in this paper were obtained under an institutional access protocol that prohibits redistribution to third parties. The data are not publicly available and cannot be shared as part of the replication package. Researchers interested in accessing the data may contact Dr. Thomas Knight (Department Chair and Instructional Professor) at the University of Florida. The replication package includes all code used to produce the results reported in the paper, as well as the authors' output files corresponding to all tables and figures. These outputs were reviewed by the Data Editor as part of the pre-publication verification process. The authors certify that they had legitimate access to the data and the rights to use them for this research.

## References


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

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

Bagues, Manuel F., & Esteve-Volart, Berta (2010). Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment. *Review of Economic Studies*, [ISSN: 0034-6527] *77*(4), 1301–1328. http://dx.doi.org/10.1111/j.1467-937X.2009.00601.x, arXiv:https://academic.oup.com/restud/article-pdf/77/4/1301/18359768/77-4-1301.pdf.

Bagues, Manuel, Sylos-Labini, Mauro, & Zinovyeva, Natalia (2017). Does the gender composition of scientific committees matter? *American Economic Review*, *107*(4), 1207–1238. http://dx.doi.org/10.1257/aer.20151211, URL https://www.aeaweb.org/articles?id=10.1257/aer.20151211.

Beaman, Lori, Keleher, Niall, & Magruder, Jeremy (2018). Do job networks disadvantage women? Evidence from a recruitment experiment in Malawi. *Journal of Labor Economics*, *36*(1), 121–157. http://dx.doi.org/10.1086/693869, arXiv:https://doi.org/10.1086/693869.

Bertrand, Marianne, & Duflo, Esther (2017). Field experiments on discrimination. *vol. 1*, In *Handbook of economic field experiments* (pp. 309–393). Elsevier.

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

Bordalo, Pedro, Coffman, Katherine, Gennaioli, Nicola, & Shleifer, Andrei (2016). Stereotypes*. *The Quarterly Journal of Economics*, [ISSN: 0033-5533] *131*(4), 1753–1794. http://dx.doi.org/10.1093/qje/qjw029, arXiv:https://academic.oup.com/qje/article-pdf/131/4/1753/30636967/qjw029.pdf.

Bordalo, Pedro, Coffman, Katherine, Gennaioli, Nicola, & Shleifer, Andrei (2019). Beliefs about gender. *American Economic Review*, *109*(3), 739–773. http://dx.doi.org/10.1257/aer.20170007, URL https://www.aeaweb.org/articles?id=10.1257/aer.20170007.

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

Card, David, DellaVigna, Stefano, Funk, Patricia, & Iriberri, Nagore (2020). Gender differences in peer recognition by economists. Working Paper.

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

Carrell, Scott E., Figlio, David N., & Lusher, Lester R. (2022). Clubs and networks in economics reviewing. http://dx.doi.org/10.3386/w29631, Working Paper 29631, National Bureau of Economic Research, URL http://www.nber.org/papers/w29631.

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

Chari, Anusha, & Goldsmith-Pinkham, Paul (2017). Gender representation in economics across topics and time: Evidence from the NBER summer institute. http://dx.doi.org/10.3386/w23953, Working Paper 23953, National Bureau of Economic Research, URL http://www.nber.org/papers/w23953.

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

Coffman, Katherine Baldiga (2014). Evidence on Self-Stereotyping and the Contribution of Ideas *. *The Quarterly Journal of Economics*, [ISSN: 0033-5533] *129*(4), 1625–1660. http://dx.doi.org/10.1093/qje/qju023, arXiv:https://academic.oup.com/qje/article-pdf/129/4/1625/30631661/qju023.pdf.

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

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

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

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

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

Hengel, Erin (2022). Are Women Held to Higher Standards? Evidence from Peer Review*. *The Economic Journal*, [ISSN: 0013-0133] http://dx.doi.org/10.1093/ej/ueac032, ueac032, arXiv:https://academic.oup.com/ej/advance-article-pdf/doi/10.1093/ej/ueac032/43881498/ueac032.pdf.

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

Kline, Patrick, Rose, Evan K., & Walters, Christopher R. (2022). Systemic Discrimination Among Large U.S. Employers*. *The Quarterly Journal of Economics*, [ISSN: 0033-5533] *137*(4), 1963–2036. http://dx.doi.org/10.1093/qje/qjac024, arXiv:https://academic.oup.com/qje/article-pdf/137/4/1963/51053979/qjac024.pdf.

Kline, Patrick, & Walters, Christopher (2021). Reasonable doubt: Experimental detection of job-level employment discrimination. *Econometrica*, *89*(2), 765–792. http://dx.doi.org/10.3982/ECTA17489, URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA17489.

Langan, A. Mark, Wheater, C. Philip, Shaw, Emma M., Haines, Ben J., Cullen, W. Rod, Boyle, Jennefer C., Penney, David, Oldekop, Johan A., Ashcroft, Carl, Lockey, Les, & Preziosi, Richard F. (2005). 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. http://dx.doi.org/10.1080/0260293042003243878, arXiv:https://doi.org/10.1080/0260293042003243878.

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

Lu, Ruiling, & Bol, Linda (2007). 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*, [ISSN: 1541-4914] *6*(2), 100–115.

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

Mengel, Friederike, Sauermann, Jan, & Zölitz, Ulf (2018). Gender Bias in Teaching Evaluations. *Journal of the European Economic Association*, [ISSN: 1542-4766] *17*(2), 535–566. http://dx.doi.org/10.1093/jeea/jvx057, arXiv:https://academic.oup.com/jeea/article-pdf/17/2/535/28264619/jvx057.pdf.

Milkman, K. L., Akinola, M., & Chugh, D. (2015). 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. http://dx.doi.org/10.1037/apl0000022.

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

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

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

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

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

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

Stoddard, Olga, Karpowitz, Christopher F., & Preece, Jessica (2021). Strength in numbers: A field experiment in gender, influence, and group dynamic. (Mimeo).

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

Zeltzer, Dan (2020). Gender homophily in referral networks: Consequences for the medicare physician earnings gap. *American Economic Journal: Applied Economics*, *12*(2), 169–197. http://dx.doi.org/10.1257/app.20180201, URL https://www.aeaweb.org/articles?id=10.1257/app.20180201.