Peer Evaluations: Exploring the Effect of Gender Pairs [*]

Thomas Knight[†]
Perihan O. Saygin[‡]

April 25, 2021

Abstract

Peer grading is widely used to evaluate low-stakes assignments in secondary and higher education settings. Additionally, peer evaluations are pervasive in the workplace, which affects hiring and promotion decisions. We conduct a peer grading experiment in a large, introductory course at a comprehensive research university. Peer graders are randomly assigned to evaluate several short essay assignments, and peer graders are incentivized to match the instructor-assigned grades as closely as possible while using a clear and structured rubric. We compare these peer-assigned grades to instructor-assigned grades considering two types of subscores: content and writing where the latter has more room for subjective evaluation and is less important in overall score. Our results suggest that validity of peer grading is much higher in content scores and for low performing students. We find that peer graders assign lower scores than instructors, and that female peer graders assign lower scores than their male counterparts, even when incentivized to match the instructor's grade. While female students receive higher peer grades in writing subscores, female graders do not appear to give any different grades to their female peers under such incentives. In real-world settings, because females are more likely to be evaluated by female peers due to gender segregation in educational, occupational, and sectoral choices, they may inherently face a tougher reviewer.

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

[*] We would like to thank Robert Ainsworth, Drew Atchison, Chris Curran, Scott Kostyshak, Peter Orazem, Rich Romano, Mark Rush, Zhenda Yin for thoughtful comments. We also would like to thank WATE 2021 and MEA-SOLE 2021 participants. All errors are our own. Corresponding author's email: psaygin@ufl.edu

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

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

1 Introduction

Performance evaluations serve a critical role in economic and non-economic activity. They are used in promotion and raise determinations for employees, compensation and options packages for executives, and grades and graduation for students. A poor evaluation can result in termination of employment or dismissal from an academic institution. Consequently, these evaluations shapes incentive structures in a wide variety of settings. They also raise fairness concerns, because they play a role in determining professional and academic success. These concerns are especially present when evaluations are conducted by peers.

Peer evaluations offer several benefits but also raise efficacy and fairness concerns. For instance, they may reduce the cost of executing the evaluation process and hasten the delivery of feedback by distributing the evaluative effort across a wider set of individuals. Additionally, peers might have better information about the tasks being evaluated and the amount and intensity of exerted effort, which is less observable by a manager or supervisor. Despite these benefits, peer evaluators often have less training and may be more subjected to explicit and implicit biases. We explore these peer evaluations in an academic setting.

Peer grading is a widely adopted practice across many different educational settings. This practice involves distributing assignment submissions to students' classmates, who then evaluate and potentially assign a grade to the work. Answer keys, grading rubrics, and clear instructions are often distributed to peer graders, which supports the efficacy of the peer review process and partially removes the potential for subjective and unfair evaluations. This peer grading procedure has been adopted in and adapted to K-12 and university classes. It is used in traditional, brick-and-mortar courses, as well as large, online courses.

There are several reasons why instructors implement peer grading in their courses. The most obvious reason is that it reduces their grading burden. Instructors who teach very large courses or several sections, have limited access to teaching assistant support, or must dedicate substantial time to research activities can assign pedagogically valuable assessments for which careful evaluation is simply not feasible without peer grading. These instructors can enhance their courses by assigning homework and papers that promote rigor and student learning. Another reason for implementing peer grading is to encourage students' thoughtful interaction with an answer sheet after an assignment's submission deadline. While distributing answers makes them available for students to review, required peer grading forces students to actively engage with those answers. Peer graders must review the correct answers to a question before evaluating their classmate's submission. This active engagement promotes learning. Finally, students might exert more effort when completing assignments if they know that their classmates will be reviewing it. This reputational effect of peer grading, however, is only likely to be present in smaller, more intimate classroom settings.

Some instructors incorporate peer-assigned grades into official assignment grades and final course grade calculations. This raises important questions about the efficacy of the peer grading process. Instructors, students, and education researchers rightly ask whether peer-assigned grades match instructor-assigned grades. Even when instructors distribute answer keys, grading rubrics, and clear instructions to peer graders, peer-assigned grades might systematically differ from instructor-assigned grades. Peer graders receive little or no instructional training, they may not have proper incentives to carefully and thoughtfully evaluate classmates' work, and they may be more likely to act on their implicit gender, racial, and ethnic biases.

Validity of peer grading and testing for potential biases is a policy-relevant issue for pedagogical purposes. Exploring peer grading can, however, teach us also about the gender bias in labor market outcomes which may be driven by peer evaluations. Hiring and promotion decisions are often made based on peer reviews/interviews and many firms rely on peer referrals as they reduce the cost of hiring. There is a vast literature on gender bias in performance evaluations and referrals and how this bias affects the gender gap in labor market outcomes. Using content analysis of individual annual performance reviews, Cecchi-Dimeglio [2017] shows that women were 1.4 times more likely to receive critical subjective feedback (as opposed to either positive feedback or critical objective feedback) data also revealed that women got less constructively critical feedback. Also, Correll and Simard [2016] suggest that women receive vaguer feedback than men do. As for the referrals, Beaman et al. [2018], Sarsons [2017a] and Zeltzer [2020] find consistent evidence that women receive fewer referrals in labor markets.

Another common application of peer evaluations is referee reports that are used to evaluate academic papers for publication. Referee reports play a central role in publication process which in turn determine academic labor market outcomes. There is growing evidence showing women academics are set a higher bar for acceptance for publications [Abrevaya and Hamermesh, 2012, Card et al., 2020b, Hengel, 2018]. This literature finds also gender differences in the evaluation of scientific work other than publication process, including studies by Card et al. [2020a] on peer recognition, Hengel [2019] on citations, Sarsons [2017b] on coauthorship and promotions, Chari and Goldsmith-Pinkham [2017] and Hospido and Sanz [2019] on conference submissions.

Our paper provides evidence on the peer reviews pre-labor market entry which contributes to understanding the sources of the biases in labor market outcomes. Previous literature on peer grading find inconclusive evidence on the validity of peer grading while there is little to no evidence on the gender bias in peer grading. Most important reason for the inconclusive results is that the validity depends on several factors related to peer grading setup which seem to differ across the existing studies. Falchikov and Goldfinch [2000] provide a meta-analysis of 56 studies conducted between 1959 and 1999 and find a 0.69 correlation between student-assigned scores and instructor-assigned scores.

Analyzing the peer grading of a sample of university students, Cho et al. [2006] suggest that clear instructions and rubric make peer reviews closer to teacher-assigned grades. Sadler and Good [2006] analyzed self-grading as well as peer grading by middle school students and found a 91-94% correlation with teacher-assigned grades with some bias with respect to achievement where high performers received lower peer grades than teachers assigned grade and low performers gave better grades to themselves. While they did not find any support for learning effect of peer grading, they found that self-grading was positively associated with learning experience. Luo et al. [2014] analyze data from Massive Open Online Courses (MOOC) and suggest that peer grades were fairly similar to teacher-assigned grades on average and they find that peer grading improved student learning.

To our knowledge, there is only a couple of studies on gender bias in peer grading. Sonnert [1995] does not find any gender bias in biologist evaluating their peer scientist's work while Langan et al. [2005] find that male undergraduate student favor their male peers when they evaluate their peers' presentations.

Varying findings on validity of peer grading can be due to several factors that differ across studies. For example, degree of clarity in instructions and rubric can affect how much the peer grades deviate from instructor-assigned grades. Also, if peer grades are used for students' actual grades, this may switch on a competition channel as well as a difficulty of giving a negative feedback to their friends[1]. This familiarity effect will also depend on whether students know who their peer grader are and if they receive their peer grades at all. We set up our experiment in a way that we can rule out several of these factors while focusing on the gender differences. The most relevant feature of our experiment is that students do not know who their peer graders are and they only receive their instructor-assigned grades. More importantly, our peer graders are given incentives to make an accurate peer assessment where their peer grading is graded based on how closely they followed the rubric.

We analyze the results of a peer grading scheme in large, mostly online course and compare peer-assigned and instructor-assigned grades for short essay assignments where overall grades are calculated as a combination of content and writing scores. We identify differences between the instructor-assigned and peer-assigned grades, which supports validity concerns. The validity concerns are stronger for writing scores compared to content scores which are based on a more deterministic and objective rubric. We also find that validity seems to be lower for high performing students. We observe that peer graders assign lower scores than instructors, and this is mostly driven by the content scores rather than the writing scores. As for the gender differences, we find evidence that female peer graders assign lower scores than male peer graders while there is no evidence for discriminating against women or men. Female students seem to have a small advantage due to peer grading in terms of the writing scores. This seems to be driven by male graders.

These findings do not imply a gender bias as the peer graders were randomly assigned and probability of having a female peer grader was not systematically different for students. In real world, however, females select into courses or jobs/occupation where they are more likely to be evaluated by females due to gender-related segregation[2]. This may then contribute to the hiring, promotion, and wage gap in sectors where peer evaluation is a common practice.

The rest of the paper's organization is as follows: sections 2 and 3 details the experimental

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

[2] In academic publishing, editors are found to be more likely to select reviewers of the same gender. (See Card et al. [2020b]

setting and our data as well as sample selection with descriptive statistics, respectively. Section 4 describes our methodology, defines our variables, and discusses our assumptions. We conclude in Section 5 discussing our main findings and potential extensions of this experiment for further research.

2 Experimental Setting

We conducted a peer grading experiment in an introductory economics course at a large, comprehensive research university. Students completed four short essay assignments during the course, which were then evaluated by a randomly matched classmate. Each assignment was based on a clear prompt to which there were objectively correct answers. No outside research was required for completing these essays. We compare the peer-assigned grades to official grades of the same assignments and explore whether these two scores systematically differ.

The experiment was conducted during the Fall 2018 and Spring 2019 semesters in a large, mostly online introductory macroeconomics course. The combined enrollment between the two semesters was 2,190 students. These undergraduate students came from all undergraduate degree-granting constituent colleges of the university and represented a wide variety of majors. Enrolled students were 46.48% male and 53.52% female.

All enrolled students were required to complete four short essay assignments. These essay assignments asked students to answer specific questions about an economic graph. There was a single objectively correct answer to each question. No outside research was required, and neither subjective analysis nor students' own opinions were solicited. These questions were the types of questions that an instructor would 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 score, 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 150-250 words.

Each of the seven short writing assignments was based on a multi-part analytical prompt for which there was a single objectively correct answer. The submissions were graded according to a rubric that included content and writing questions, generating 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 score (defined as the percentage of total possible points earned). The same rubric was used by both expert graders and peer graders, and it was released shortly after the submission deadline.

Assignments are submitted into an electronic course management system (Canvas). Each assignment submission was graded by trained graduate teaching assistant, who assigned the official score for inclusion in the student's final course grade calculation, and by a randomly assigned peer grader for evaluation. Both teaching assistants and peer graders evaluated the assignments using a common rubric. Rubrics contained two types of questions: between seven and eleven economic content questions, 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 was scores on the zero-to-ten scale.

When completing their evaluations, both teaching assistants and peer graders had access to the same information and grading rubrics. They could view the student's first and last name, a small picture of the student's face, and the essay submission. They were instructed to read the submission and complete the grading rubric by scoring each question on the rubric on the zero-to-ten scale. Peer graders had one week to complete the peer review. Official expert grades were released after one week while teaching assistants did not have access to peer grades.

The official scores that were used to calculate final course grades were the teaching assistant-assigned grades. Each writing assignment accounted for 2% of the authors' final course grade. Peer grades only affected the final course grade of the reviewer. Peer graders were told that they must complete a peer review and that it must "more or less" match the the expert graders' evaluations to receive credit. Peer graders received an overall peer grading score (across all assignments) that accounts for 4% of their final course grade.

In order to detect or identify the gender bias in peer grading, it is important to highlight the most relevant features of the peer grading experiment. In our setup, peer-assigned grades do not affect the student's final course grade. Studying the gender bias when peer reviewers know that their evaluations affect someone else's grade could provide important insights, because that is what often happens in "real-world" settings. However, if the peer grades were used in determining final course grades, rather than instructor or teaching assistant–assigned grades, it would become challenging to isolate potential several channels that may be at play. In our setup, for example, we do not expect to observe any competition or familiarity effect, which could also generate systematic deviations from the instructor-assigned grades. Last but not least, peer graders' final course grades are affected by their careful and thoughtful completion of peer reviews. This gives them incentives to be as precise as possible. This feature allows us to see if there is a gender bias when they are incentivized to be precise. Discrimination literature suggests that the observed discrimination tends to be lower or diminish when there is explicit or implicit monitoring (See for example, Parsons et al. [2011]). Similarly, under these incentives of our experiment, we expect to identify a lower bound for the gender bias in peer grading, if any.

3 Sample Selection and Summary Statistics

The data are generated by the teaching assistants' expert evaluations and classmates' peer reviews of seven short writing assignments. After the submission deadline, each student's assignment was added to the teaching assistants' grading queue and then randomly assigned to one of the student's classmates for peer review. The peer review assignments were made randomly by the course management system, and the matching process was based solely on the pool of students who submitted the assignment by the deadline. If a student did not submit their assignment by the deadline, they received a zero for their own assignment and were not assigned a peer review, which implied a zero on that peer review exercise as well. Each students 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 are ever enrolled in the course in Fall 2018 and Spring 2019. On the other hand, in our peer grading analysis, inclusion in the sample is conditional on submitting the assignment by the deadline and completing the course. Students who failed to submit a particular assignment by the submission deadline are not included in the data for that assignment. These students will, however, appear in the data for other assignments. Also, we have missing instructor-assigned or peer-assigned grades if students fail to follow instructions and/or if a peer grader does not complete the peer grading. Ultimately, we work with the sample of homework assignments that are submitted correctly and if their assigned peer completed the peer grading.

In Table 1, we report the descriptive statistics for this selection on the full sample of students for 7 assignments over the two semesters. 93% of students completed the course on average and 92% of students completed at least one peer grading. Approximately 10% of assignments were submitted incorrectly and Among assigned peer graders 1% of them did not complete their peer review. In terms of incorrect submission and incomplete peer grading, we do not find any significant differences between men and women. On the other hand, 11% of the assignments were not submitted at all on average and male students are more likely to skip an assignment which seems to be the only significant gender difference.

Descriptive statistics of performance measures are reported for the full sample in Table 2. We report the same statistics for the sample of our analysis in Table 3. Female students perform better than male students in all assessments except for the exams. This finding is consistent with the previous findings on gender differences in performance under pressure and when stakes are high. There seems to be small differences in these statistics in the sample of our analysis compared to the full sample where our sample is slightly more positively selected but the homework assignments and peer scores and the gender differences seems to be similar.

Our summary statistics in Table 3 shows that there is a small performance gap in favor of female students in overall assignment scores assigned by teaching assistants. This gap is largely driven by content score while the difference between male and female students' TA-assigned writing subscores is relatively small and hardly significant. When we look at the peer-assigned grades, the gender gap is larger but it has the same pattern in content and writing scores as in teaching assistant-assigned grades. The peer assigned grade averages seem to be smaller than TA-assigned averages both for female and male students.

Lastly, we check for balance of characteristics between the assignments graded by female and male peers. In Table 4, we show the summary statistics of peer assigned scores by the gender of the peer grader. It seems that peer graders were not more or less likely to be assigned to a student with a certain gender in our randomized peer assignment. Also, female and male peer graders seem to be assigned to students who performed similarly in terms of 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 largest in content subscores. Based on the summary statistics, it seems that peer graders are tougher graders than the teaching assistants while female peers seem to deviate negatively even more than male peer peer graders. In our analysis below, we will investigate these differences in further details.

4 Methodology and Results

Our empirical model is a reduced form model that estimates the overall gender differences in peer scores depending on the gender of the student and the peer grader. For a student i in assignment j, the model is given by:

$$\text{Peer Score}_{ij} = \alpha + \beta_1 FS_i + \beta_2 FPG_{ij} + \beta_3 (FS_i \times FPG_{ij}) + \beta_4 \text{Score}_{ij} + \mu_j + \epsilon_{ij} \qquad (1)$$

where Peer Score$_{ij}$ is the peer-assigned score and Score$_i j$ is the TA-assigned score of student i in assignment j and ϵ_{ij} is a random error term. FS_i is an indicator variable taking value 1 if the student i is female, FPG_{ij} is an indicator variable taking value 1 if student i is assigned to a female peer grader in assignment j, and μ_j captures the fixed effects for assignments. Standard errors are clustered at the student level and the hypothesis to be tested are $\beta_1 \neq 0$ in order to check whether female students have an advantage in peer grading, and $\beta_2 \neq 0$ in order to see whether scores assigned by female peer graders are systematically different from those assigned by male peer graders. In the next step, we analyze whether there is a differential effect of being assigned to a female student for female students and test whether $\beta_3 \neq 0$ to test this hypothesis.

4.1 Results: Gender Differences

We first analyze the overall peer scores in Table 5. The first column shows that female students' peer-assigned scores are 1.53 percentage points higher than male students and this difference is statistically significant. Also, peer scores seem to be lower if they are assigned by female peer grader by 1.79 percentage point. In the second column, we add control for the TA-assigned score. The female advantage becomes much smaller and it is only marginally significant. While we observe a reduction in the coefficient of female grader, we still find that female peer graders assign 1.31 percentage points lower scores than their male counterparts on average conditional on the TA-assigned scores. In the third column, we show that there is no evidence for a differential effect of being assigned to a female peer grader by the gender of students. It is important to note that peer graders are incentivized to match the rubric as closely as possible which creates a monitoring mechanism and such mechanisms tend to lower the potential biases. In column 4 and 5, we explore the possibility that peer-assigned scores may be affected by peers' own performance on the same assignment or knowledge of the material. We find that conditioning on graders' own score on the same assignment does not change our results. Columns 4 and 5 also confirm further that we do not observe any differential effects of being assigned to a female peer for female (or male) students.

In the next step, we analyze the subscores in a similar fashion in Table 6 and Table 7. While the sign and significance of our variable of interests remain mostly consistent with the analysis of the overall scores, we find somewhat different results in content subscores compared to the writing subscores. In Table 6, we find that peer grading does not favor female students' peer content scores conditional on the TA-assigned content scores while assignments graded by female graders are found to receive 1.55 points lower content scores conditional on the TA-assigned content scores which is robust to inclusion of graders' own TA-assigned scores. Table 7 shows that female students have an advantage in peer writing scores even conditional on TA-assigned writing scores. Also, female peer graders give 0.69 points lower writing scores which is smaller than the coefficient obtained in content score estimations.

These findings suggest that female peer graders are tougher graders than male peer graders in particular in content scores. Female students receive higher peer writing grades even conditional on the TA-assigned writing scores. Similar to overall scores, interaction term is insignificant in both content and writing score estimations. Peer grader's own performance do not seem to affect the results.

In order to observe whether this observation holds across the distribution of scores, we plot the cumulative distribution functions of overall scores and each subscore assigned by peer graders comparing them by the gender of peer grader. These CDFs are presented in Figure 4 and Figure 5. These figures, while mostly consistent with the regression results, imply that the differences are larger in the content scores and contents graded by male peer graders seem to be stochastically dominant over contents graded by female peer graders.

4.2 Results: Validity of Peer Grading

The same regressions in the previous section allow us to analyze the validity of peer grading which is measured by the statistical association between the peer scores and TA-assigned scores. The coefficient of TA-assigned scores are much 0.83 and 0.20 for content scores and writing scores respectively. This implies that the validity is a bigger concern for the writing scores. This finding is consistent with the literature suggesting stronger validity when the rubric is deterministic and objective. In writing scores, there is room for more subjective evaluation and this is reflected in the lower correlation between TA-assigned and peer-assigned writing scores.

In order to check whether the validity of peer grading is the same across the distribution of scores, we first plot the cumulative distribution functions of overall scores and each subscore assigned by peer graders and teaching assistants. These CDFs are presented in Figure 2 to Figure 3. These figures imply that the peer content scores are lower than TA-assigned content scores pretty much across the whole distribution and the differences are slightly larger in the higher end of the score distributions.

As a second step, we run quantile regressions of peer-assigned scores on TA-assigned scores and plot the coefficients at various quantiles in Figure 1. This figure shows that the association between the TA-assigned and peer-assigned scores are almost perfect below median while it declines substantially for higher quantiles while the OLS estimates are around 0.83. This implies that peer grading is highly valid for the low performing students. Given that the peers are tougher graders than the expert graders, it seems like it could potentially penalize the high performing students. In the next section, we investigate the deviation of peer grades from TA-assigned grades further.

4.3 Results: Deviation of Peer Grades from TA-Assigned Grades

In this section we focus on the deviation of peer grades from the TA-assigned grades. The figure 6 shows the average deviations by gender pairs for overall scores. Consistently with our previous findings, we find that peer graders deviate negatively on average. This negative deviation is larger among female graders compared to male graders but there is no evidence for bias against any gender. In Table 8, we report the average deviations in overall, content, and writing scores. On average they are all negative and female graders negative deviation is the largest for content subscores. In the second part of the table, we report the share of peer grades that matched the TA-assigned grades, and the share of those who deviated negatively and positively. We find that more than half of peer grades matched the TA-assigned grades both in content and writing subscores. While the share of positive and negative deviations seem to be balanced for writing scores, the share of negative deviations are much larger than the share of positive deviations in content scores. Also, while there is no evidence of gender differences in these shares for writing scores, the share of negatives is higher and the share of positive is lower among female peer graders compared to male peer graders.

Finally, we focus on the deviation of the peer-assigned scores from TA-assigned scores as our dependent variable in Table 9. We find that the deviation is 1.3 points smaller in female-peer-assigned scores. We follow the same steps in our estimation specifications by controlling for TA-assigned scores and peer graders' own scores and we find consistent estimates with our previous findings. We run the same analysis for the content and writing subscores in Table 10 and Table 11.

Based on these findings combined with the previous sections, we argue that while peer graders are more likely to assign lower scores than the teaching assistants, female peers are even tougher graders than their male counterparts. This gender difference seems to be mostly driven by content scores rather than writing scores. We find that female students receive more favorable peer grades in writing scores while this is not the case in content scores. Last but not least, we do not find any evidence for females assigning any different scores based on the gender of the student under the incentives for staying loyal to rubric as closely as possible.

5 Conclusion

We conduct a peer evaluation experiment using peer grading in an introductory economics course at a large, comprehensive research university. Students complete short writing assignments, for which there is an objectively correct answer, and each submission is evaluated by a trained graduate teaching assistant and a randomly-assigned classmate. Overall grades are calculated as a combination of content and writing scores 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 teaching assistant-assigned grades in order to receive credit.

We compare teaching assistant-assigned grades and peer-assigned grades to evaluate the general efficacy 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 efficacious if peer grades match the the grades assigned by trained teaching assistants. 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 systematically differ. The validity concerns are stronger for writing scores compared to content scores which are based on a more deterministic and objective rubric. We also find that validity seems to be lower for high performing students.

Second, we observe that peer graders are tougher than expert graders, and this is mostly driven by the content scores rather than the writing scores. Female peer graders assign overall scores that are 2.16 percentage points lower than the teaching assistants, and male peer graders assign overall scores that are 0.86 percentage points lower than the teaching assistants. While these may appear to be small deviations, they are statistically significant and can be interpreted as a lower bound. Because peer graders were incentivized to match the experts' scores, the observed deviation is potentially smaller than it would be without this "monitoring". Monitoring has been shown to reduce biases and result in more accurate peer evaluations.

One might be concerned that teaching assistants may be more prone to inflating grades if they wish to avoid complaints and contested scores. This could generate the observed deviations between expert and peer grades. However, in this experimental setting, complaints and contested scores were not handled by teaching assistants, but instead by the course instructor. The teaching assistants were simply instructed to evaluate the work and assign a grade. There were no apparent incentives to inflate or deflate scores.

Third, we also find that female students are tougher peer graders than male students. Female peer graders assign overall scores that are 1.31 percentage points lower than male peer graders assign conditional on TA-assigned scores. We also considered that course performance and understanding of the material might differ by gender, on average, 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.

Finally, we do not observe any bias toward or against female students. While female students receive higher evaluations from their peers, this observation disappears after taking account of the teaching assistants' grade of the same assignment in content scores. The female advantage remains significant in writing scores even after conditioning on TA-assigned writing scores. Given that writing score has a much smaller weight, this advantage is not reflected significantly n overall scores.

While we observe no evidence of gender bias, it is possible that the experimental setting itself combats such a bias. Because peer graders are monitored and even incentivized, they may demonstrate less bias than they would in a "real world" setting without monitoring and/or incentives. That is, it may be the case that peer graders do possess gender-based biases, but that they did not act on those biases when they know they were being observed. Moreover, in an experimental setting, female students were equally likely to be assigned to female or male peer graders. In a "real world" setting where females make tougher evaluations, female students and workers will be more likely to be receiving lower grades or performance evaluations since they will be more likely to share a workplace or a classroom with other female peers due to gender segregation in labor markets and educational decisions.

References

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

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

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

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

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

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

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

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

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

Erin Hengel. Publishing while female. are women held to higher standards? evidence from peer review. *Working Paper*, 2018.

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

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

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

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

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

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

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

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

Heather Sarsons. Gender differences in recognition for group work. *Working Paper*, 2017b.

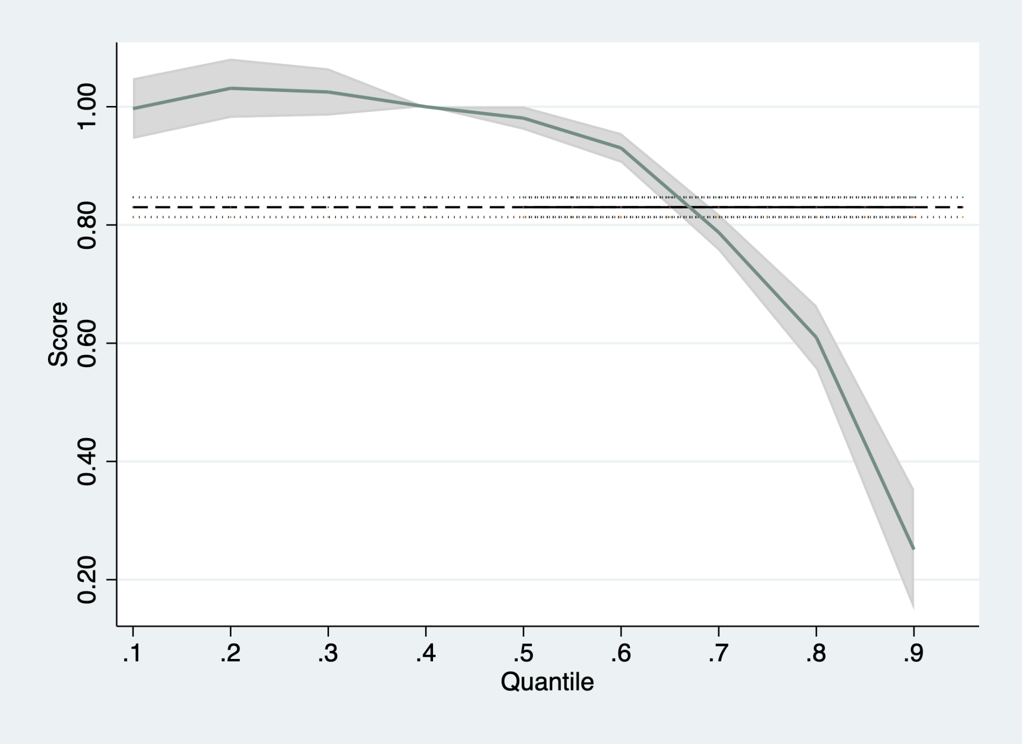
Gerhard Sonnert. What makes a good scientist?: Determinants of peer evaluation among biologists. *Social Studies of Science*, 25(1):35–55, 1995. doi: 10.1177/030631295025001003. URL `https://doi.org/10.1177/030631295025001003`.

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

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

6 Figures and Tables

Figure 1: Peer Score Quantile Regression: Coefficients of Score



Note: Coefficients of instructor-assigned scores obtained from a quantile regression are plotted together with the OLS coefficient and confidence intervals.

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

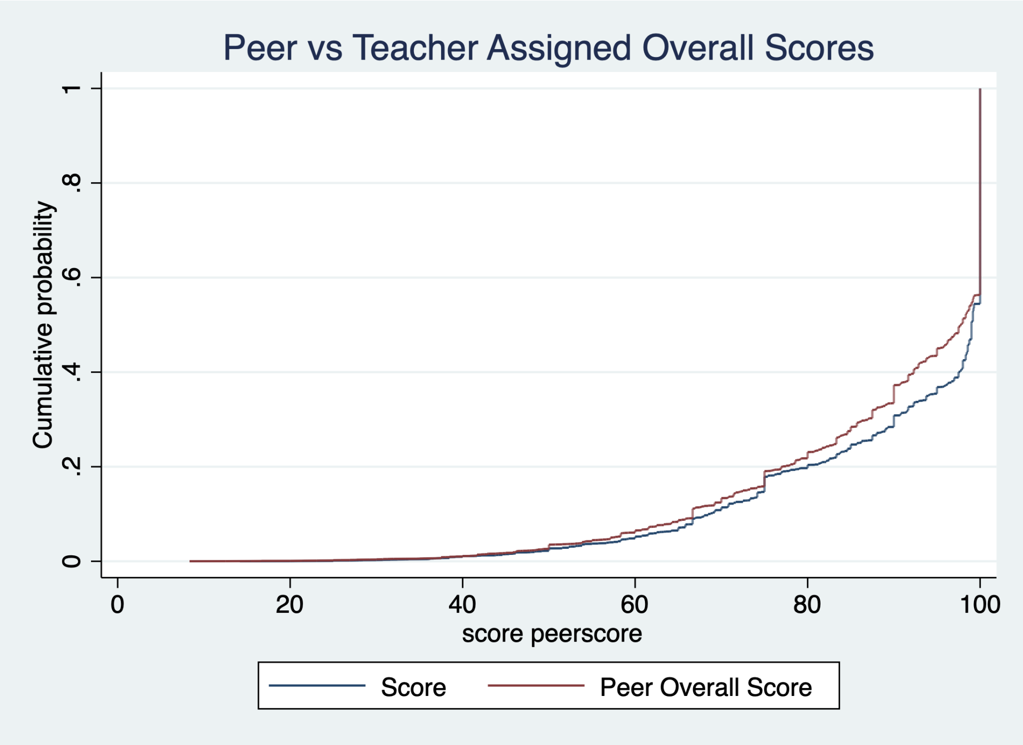
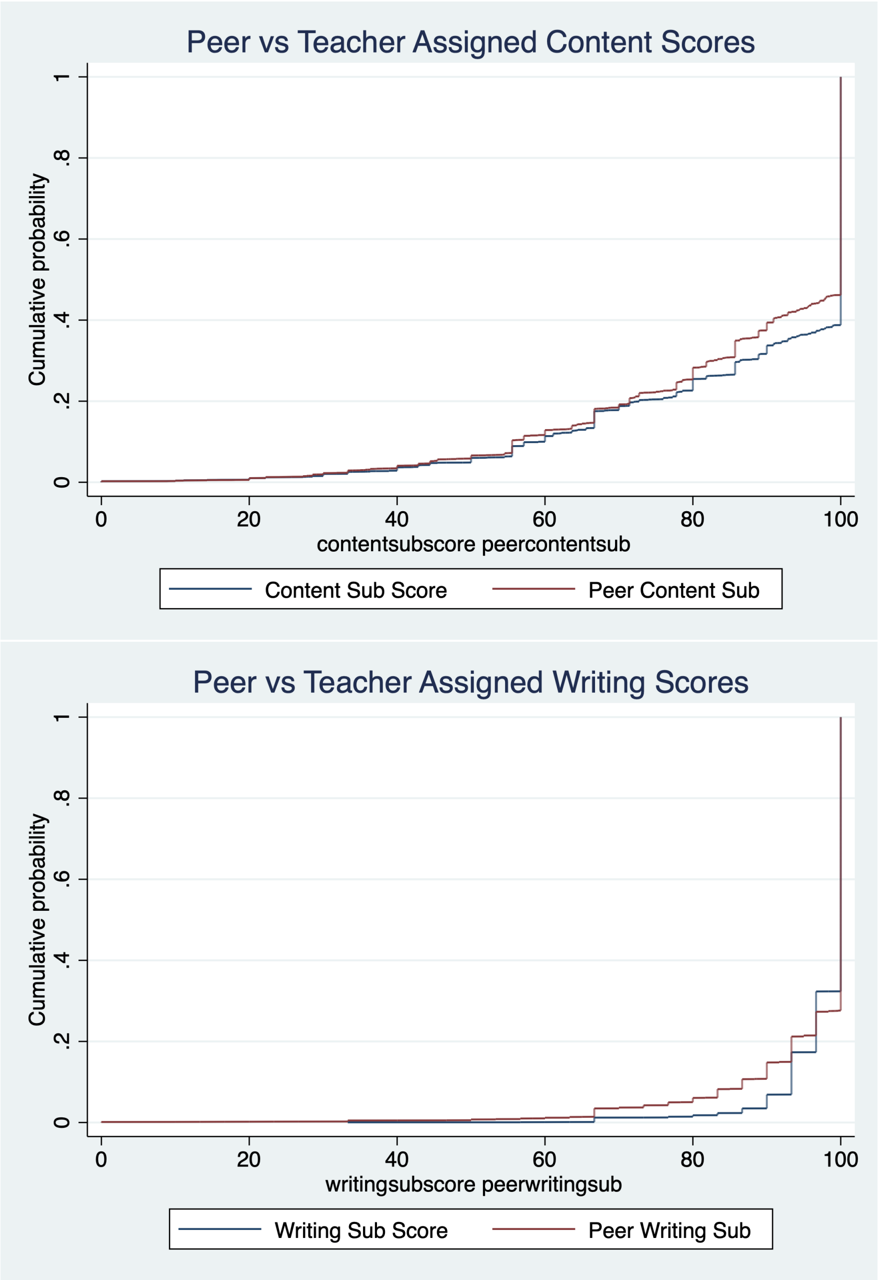


Figure 3: CDF of Subcores Assigned by Instructors and Peers



Peer vs Teacher Assigned Content Scores

Peer vs Teacher Assigned Writing Scores

Figure 4: CDF of Peer-assigned Grades by Peer Grader's Gender

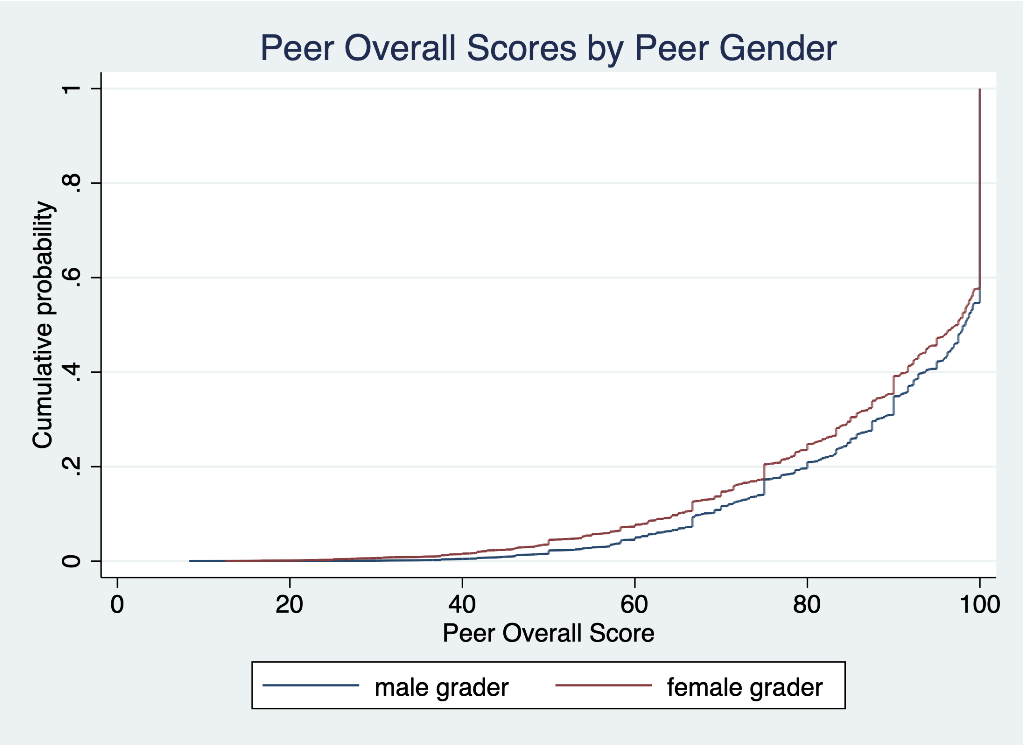
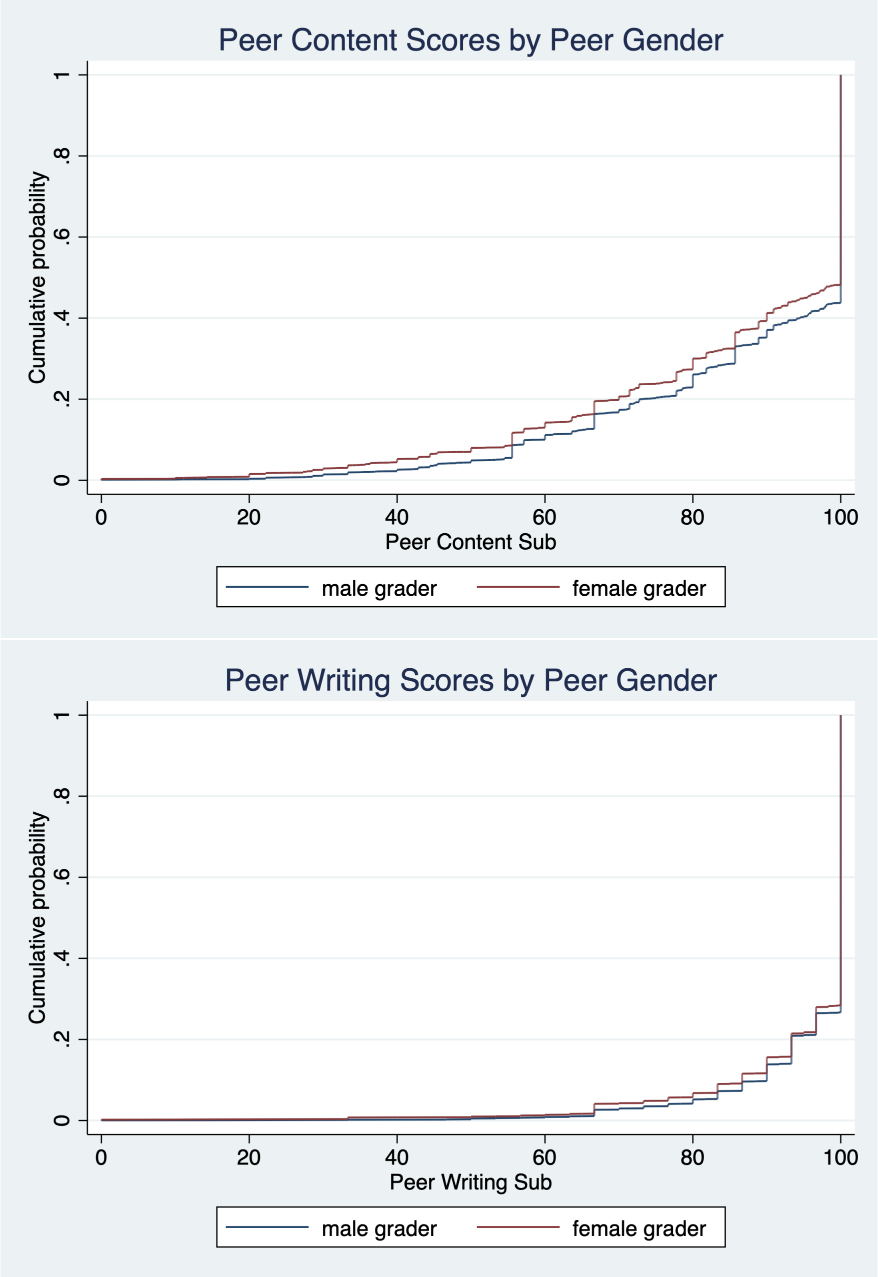


Figure 5: CDF of Subscores Assigned by Peers by Peer Grader's Gender



Peer Content Scores by Peer Gender

Peer Writing Scores by Peer Gender

Figure 6: Deviations of peer graders by peer graders' gender

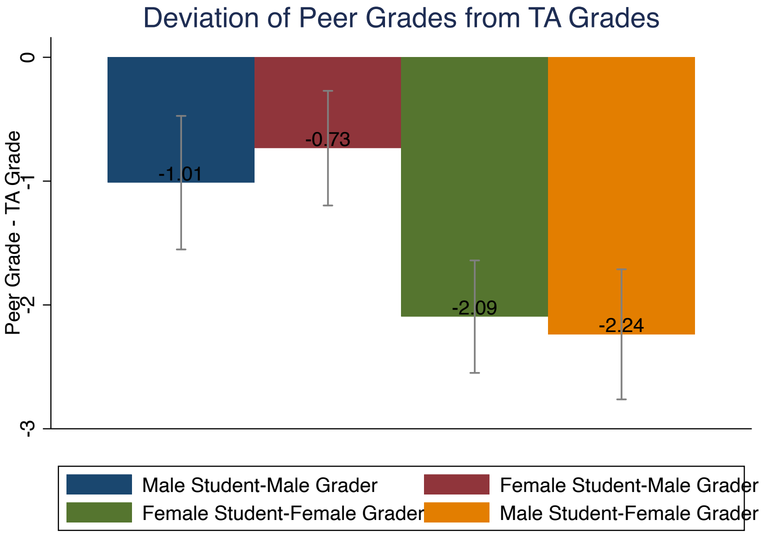


Table 1: Summary Statistics by Gender (All): Selection

	Female	Male	Gap
	Mean/sd	Mean/sd	b
Received a Final Course Grade	0.93	0.93	-0.00
	(0.26)	(0.26)	
Received a Peer Review Score	0.93	0.92	-0.00
	(0.26)	(0.27)	
No homework submitted	0.10	0.12	0.02**
	(0.30)	(0.33)	
Peer assigned but no peer grade turned in	0.01	0.01	-0.00
	(0.12)	(0.11)	
Homework submitted incorrectly	0.09	0.10	0.01
	(0.29)	(0.30)	
Homework Submitted in TA bin but not in Peer Bin	0.08	0.09	0.01
	(0.27)	(0.29)	
Homework Submitted in Peer Bin but not in TA Bin	0.01	0.01	0.00
	(0.10)	(0.10)	
Observations	4056	3507	7563

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

Table 2: Summary Statistics by Gender (All): Performance

	Female Mean/sd	Male Mean/sd	Gap b
Exam 1 (3784768)	81.13	83.37	2.24***
	(14.09)	(12.73)	
Exam 2 (3801850)	75.59	76.65	1.06**
	(15.14)	(13.38)	
Exam 3 (3828226)	81.42	84.45	3.03***
	(15.34)	(13.35)	
Score	91.04	89.95	-1.09**
	(14.38)	(15.23)	
Content Sub Score	88.64	87.30	-1.34**
	(18.93)	(20.08)	
Writing Sub Score	97.68	97.34	-0.34*
	(5.10)	(5.70)	
Peer Overall Score	89.84	88.41	-1.43***
	(14.87)	(15.72)	
Peer Content Sub	87.44	85.74	-1.70***
	(19.24)	(20.27)	
Peer Writing Sub	96.47	95.78	-0.68**
	(8.90)	(10.22)	
Final Quiz Average (3736450)	91.78	90.17	-1.61***
	(10.85)	(12.85)	
Observations	4023	3485	7508

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

Table 3: Summary Statistics by Gender (Analysis Sample): Performance

	Female	Male	Gap
	Mean/sd	Mean/sd	b
Exam 1 (3784768)	82.96	84.87	1.90***
	(12.71)	(11.36)	
Exam 2 (3801850)	77.16	77.72	0.56
	(14.25)	(12.62)	
Exam 3 (3828226)	82.29	84.83	2.54***
	(14.65)	(13.32)	
Score	91.39	90.28	-1.11**
	(14.00)	(14.99)	
Content Sub Score	89.05	87.67	-1.38**
	(18.49)	(19.87)	
Writing Sub Score	97.80	97.51	-0.29*
	(4.88)	(5.36)	
Peer Overall Score	89.89	88.60	-1.29**
	(14.79)	(15.57)	
Peer Content Sub	87.51	86.01	-1.50**
	(19.18)	(20.01)	
Peer Writing Sub	96.48	95.81	-0.67**
	(8.87)	(10.21)	
Final Quiz Average (3736450)	92.79	91.61	-1.18***
	(8.76)	(10.04)	
Observations	3232	2694	5926

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

Table 4: Summary Statistics by Peer Grader's Gender

	Graded by Female Grader	Graded by Male Grader	Difference
	Mean/sd	Mean/sd	b
Female Student	0.55	0.53	-0.02
	(0.50)	(0.50)	
Score	90.58	91.27	0.70
	(14.87)	(13.94)	
Content Sub Score	88.00	88.95	0.95
	(19.66)	(18.46)	
Writing Sub Score	97.65	97.69	0.04
	(4.99)	(5.24)	
Exam 1 (3784768)	83.54	84.18	0.64*
	(12.34)	(11.91)	
Exam 2 (3801850)	77.39	77.44	0.05
	(13.62)	(13.43)	
Exam 3 (3828226)	83.22	83.72	0.50
	(14.21)	(13.99)	
Final Quiz Average (3736450)	92.22	92.29	0.07
	(9.47)	(9.26)	
Final Course Grade (3736448)	83.86	84.20	0.34
	(10.27)	(10.00)	
Peer Overall Score	88.42	90.41	1.99***
	(16.08)	(13.85)	
Peer Content Sub	85.73	88.19	2.46***
	(20.68)	(18.01)	
Peer Writing Sub	95.84	96.59	0.75**
	(10.40)	(8.24)	
Observations	3294	2632	5926

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

Table 5: Peer Grading: Overall Peer Scores

	(1)	(2)	(3)	(4)	(5)
Female Student	1.529***	0.514*	0.577	0.508*	0.571
	(0.449)	(0.250)	(0.349)	(0.250)	(0.349)
Female Peer Grader	-1.787***	-1.310***	-1.248***	-1.282***	-1.220**
	(0.388)	(0.241)	(0.373)	(0.242)	(0.372)
Score		0.835***	0.835***	0.835***	0.835***
		(0.013)	(0.013)	(0.013)	(0.013)
Female * Female Grader			-0.114		-0.115
			(0.487)		(0.487)
Graders Score				-0.027**	-0.027**
				(0.009)	(0.009)
Observations	5773	5773	5773	5773	5773

Note: Standard errors are in parentheses and clustered at student level. The dependent variable is overall peer-assigned scores. All estimations control for fixed effects for assignments. * p<0.10, ** p<0.05, *** p<0.010

Table 6: Peer Grading: Content Peer Scores

	(1)	(2)	(3)	(4)	(5)
Female Student	1.794**	0.560	0.498	0.554	0.493
	(0.582)	(0.318)	(0.450)	(0.318)	(0.449)
Female Peer Grader	-2.220***	-1.546***	-1.607***	-1.523***	-1.583***
	(0.508)	(0.308)	(0.469)	(0.309)	(0.468)
Content Sub Score		0.825***	0.825***	0.825***	0.825***
		(0.014)	(0.014)	(0.014)	(0.014)
Female * Female Grader			0.112		0.111
			(0.620)		(0.619)
Graders Content Sub				-0.018*	-0.018*
				(0.009)	(0.009)
Observations	5773	5773	5773	5773	5773

Note: Standard errors are in parentheses and clustered at student level. The dependent variable is peer-assigned content subscores. All estimations control for fixed effects for assignments. * p<0.10, ** p<0.05, *** p<0.010

Table 7: Peer Grading: Writing Peer Scores

	(1)	(2)	(3)	(4)	(5)
Female Student	0.737**	0.671**	1.003**	0.670**	1.003**
	(0.254)	(0.250)	(0.333)	(0.250)	(0.333)
Female Peer Grader	-0.691**	-0.695**	-0.369	-0.690**	-0.362
	(0.244)	(0.243)	(0.392)	(0.242)	(0.391)
Writing Sub Score		0.197***	0.197***	0.197***	0.197***
		(0.037)	(0.037)	(0.037)	(0.037)
Female * Female Grader			-0.602		-0.604
			(0.495)		(0.495)
Graders Writing Sub				-0.022	-0.022
				(0.024)	(0.024)
Observations	5773	5773	5773	5773	5773

Note: Standard errors are in parentheses and clustered at student level. The dependent variable is peer-assigned writing subscores. All estimations control for fixed effects for assignments. * p<0.10, ** p<0.05, *** p<0.010

Table 8: Summary Statistics of Deviation of Peer Scores from Instructor Assigned Grades by Peer Grader's Gender

	Female Grader	Male Grader	Difference
	Mean/sd	Mean/sd	b
Peer Score Deviation	-2.16	-0.86	1.29***
	(10.06)	(9.22)	
Peer Content Score Deviation	-2.27	-0.76	1.51***
	(12.78)	(11.86)	
Peer Writing Score Deviation	-1.81	-1.10	0.71**
	(10.84)	(9.30)	
Exact Match with Content Score	0.56	0.58	0.01
	(0.50)	(0.49)	
Negative Deviation in Content Score	0.30	0.26	-0.03**
	(0.46)	(0.44)	
Positive Deviation in Content Score	0.14	0.16	0.02*
	(0.35)	(0.37)	
Exact Match with Writing Score	0.53	0.53	0.01
	(0.50)	(0.50)	
Negative Deviation in Writing Score	0.24	0.23	-0.01
	(0.43)	(0.42)	
Positive Deviation in Writing Score	0.23	0.24	0.00
	(0.42)	(0.42)	
Observations	3294	2632	5926

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

Table 9: Peer - TA Difference: Overall Scores

	(1)	(2)	(3)	(4)	(5)	(6)
Female Student	0.313	0.358	0.514*	0.577	0.508*	0.571
	(0.254)	(0.363)	(0.250)	(0.349)	(0.250)	(0.349)
Female Peer Grader	-1.215***	-1.171**	-1.310***	-1.248***	-1.282***	-1.220**
	(0.249)	(0.383)	(0.241)	(0.373)	(0.242)	(0.372)
Female * Female Grader		-0.081		-0.114		-0.115
		(0.502)		(0.487)		(0.487)
Score			-0.165***	-0.165***	-0.165***	-0.165***
			(0.013)	(0.013)	(0.013)	(0.013)
Graders Score					-0.027**	-0.027**
					(0.009)	(0.009)
Observations	5773	5773	5773	5773	5773	5773

Note: Standard errors are in parentheses and clustered at student level. The dependent variable is the deviation of overall peer-assigned scores from the overall instructor-assigned grades. All estimations control for fixed effects for assignments. * p<0.10, ** p<0.05, *** p<0.010

Table 10: Peer - TA Difference: Content Scores

	(1)	(2)	(3)	(4)	(5)	(6)
Female Student	0.297	0.204	0.560	0.498	0.554	0.493
	(0.323)	(0.470)	(0.318)	(0.450)	(0.318)	(0.449)
Female Peer Grader	-1.402***	-1.494**	-1.546***	-1.607***	-1.523***	-1.583***
	(0.319)	(0.485)	(0.308)	(0.469)	(0.309)	(0.468)
Female * Female Grader		0.169		0.112		0.111
		(0.641)		(0.620)		(0.619)
Content Sub Score			-0.175***	-0.175***	-0.175***	-0.175***
			(0.014)	(0.014)	(0.014)	(0.014)
Graders Content Sub					-0.018*	-0.018*
					(0.009)	(0.009)
Observations	5773	5773	5773	5773	5773	5773

Note: Standard errors are in parentheses and clustered at student level. The dependent variable is the deviation of overall peer-assigned content scores from the instructor-assigned content scores. All estimations control for fixed effects for assignments. * p<0.10, ** p<0.05, *** p<0.010

Table 11: Peer - TA Difference: Writing Scores

	(1)	(2)	(3)	(4)	(5)	(6)
Female Student	0.399	0.810*	0.671**	1.003**	0.670**	1.003**
	(0.271)	(0.381)	(0.250)	(0.333)	(0.250)	(0.333)
Female Peer Grader	-0.712**	-0.309	-0.695**	-0.369	-0.690**	-0.362
	(0.267)	(0.430)	(0.243)	(0.392)	(0.242)	(0.391)
Female * Female Grader		-0.743		-0.602		-0.604
		(0.544)		(0.495)		(0.495)
Writing Sub Score			-0.803***	-0.803***	-0.803***	-0.803***
			(0.037)	(0.037)	(0.037)	(0.037)
Graders Writing Sub					-0.022	-0.022
					(0.024)	(0.024)
Observations	5773	5773	5773	5773	5773	5773

Note: Standard errors are in parentheses and clustered at student level. The dependent variable is the deviation of overall peer-assigned writing scores from the instructor-assigned writing scores. All estimations control for fixed effects for assignments. * p<0.10, ** p<0.05, *** p<0.010