Departament d'Economia Aplicada
Universitat Autònoma de Barcelona

WORKING PAPER 25-10

Gender Biased Resistance to Harsh Feedback

Perihan O. Saygin
Garrison Pollard
Thomas Knight
Mark Rush

Departament d'Economia Aplicada
Edifici B
Universitat Autònoma de Barcelona
08193 Bellaterra, Barcelona, Spain.
d.econ.aplicada@uab.cat
www.uab.cat/en/applied-economics
@ecapuab.bsky.social

UAB
Universitat Autònoma de Barcelona

Gender Biased Resistance to Harsh Feedback*

Perihan O. Saygin† Garrison Pollard‡ Thomas Knight§ Mark Rush¶

September 3, 2025

Abstract

Responses to performance feedback play a critical role in shaping future outcomes in educational and professional contexts. This paper examines whether evaluator gender influences the likelihood that individuals contest feedback. Using an experiment conducted in large introductory economics courses, we exploit the random assignment of evaluators with randomly assigned male- or female-sounding names to identify a systematic gender bias: individuals are significantly more likely to contest feedback when it is delivered by an evaluator with a female-sounding name than when similar feedback comes from a male-sounding evaluator. This gender disparity is most pronounced when evaluations are harsh relative to a "fair" assessment, fall short of students' performance expectations, and are more ambiguous. These findings suggest that women in evaluative positions face disproportionate resistance when delivering negative assessments and have implications for their authority, credibility, and career advancement in both educational and workplace settings.

JEL Classification: J16, J71
Keywords: gender, backlash, stereotypes.

*This study received IRB approval from University of Florida (IRB202201907) and was pre-registered with the American Economic Association (AEARCTR-0010898). We would like to thank Scott Kostyshak for inspiring suggestions at early stages of the project. We would like thank Katie Coffman, Christine Exley, Nagore Iriberri, Tanushree Jhunjhunwala, Nolan Pope, Pedro Rey Biel, Rich Romano, and David Sappington for helpful comments. All errors are our own.

†Corresponding author: peri.saygin@uab.cat, Department of Applied Economics, Universitat Autonoma de Barcelona
‡garrisonpollard@ufl.edu, Department of Economics, University of Florida
§thomas.knight@ufl.edu, Department of Economics, University of Florida
¶markrush@ufl.edu, Department of Economics, University of Florida

1 Introduction

Individuals routinely receive performance feedback in educational and professional settings. This feedback serves to inform the individual of their own performance and provide guidance for improvement. Additionally, feedback may provide an externally visible signal of ability. For example, in educational settings, instructors provide grades to students. Students also evaluate their instructors at the end of the term. In professional settings supervisors provide routine performance evaluations. Feedback appears outside of school and work in the form of customer satisfaction surveys and product reviews. Such feedback of previous performance can affect future effort and outcomes. It may also affect the recipient's employability or eligibility for promotion and raises.

Given the significant consequences of feedback, understanding individuals' responses has important implications for education, management, and other areas. In many contexts individuals can *resist* feedback if they have the option of challenging or appealing negative feedback. Despite the clear importance of performance evaluations in many educational and workplace settings, as well as their potential to impact future decision making, relatively little is known about resistance to feedback. Our paper aims to fill this gap.

The decision to resist by appealing an evaluation may depend on the source of the feedback, as well as its *harshness*. In this paper, the goal is to test whether individuals are more likely to contest overly-harsh feedback and, most critically, examine whether this decision depends on the gender of the evaluator. In particular we test whether individuals show greater resistance to especially harsh feedback and if this resistance increases disproportionately when the evaluator is a woman.

The most substantial challenge in answering these questions is the lack of an exogenous variation in feedback. High-performing individuals tend to receive positive feedback, while low-performing individuals generally receive negative feedback. Often, the "fair" assessment is unobserved. Methodologically, the econometrician cannot identify whether an evaluation is overly harsh or overly lenient in most real-world environments. Moreover, in many educational and professional settings, there is endogenous sorting of individuals and their evaluators. It is difficult to isolate the effect of being evaluated by a particularly lenient or harsh supervisor, as the leniency of supervisors may correlate with other characteristics that affect the supervisees' responses to feedback.

Our paper overcomes these challenges in an experimental setting in introductory economics classes at a flagship public university. We review students' reactions to their scores on 8,498 essays. We randomly assign several graduate student graders to evaluate the essays. The graders and students are totally independent. The essays do not have the students' names and the graders do not interact with the students in any way other than grading the essays. Despite a random allocation of graders to essay submissions and a common scoring rubric, we observe that the average grades awarded by graders differ systematically. Some graders are inherently harsher than others when it comes to subjective evaluation of an essay's writing quality or in assigning partial credit for partially correct answers. The random assignment of these graders with varying levels of harshness generates exogenous variation in grades.[1]

While we observe an essay's assigned score, we need to quantify the harshness of the score. To obtain a measure of harshness, we constructed a performance benchmark, which we refer to as the "fair" score. After the semester concluded and grades were submitted, we trained and then randomly assigned two additional graders to evaluate and score each assignment without being informed of the initial score nor the student's name. We deem the average of these three grades—the two ex post scores and the score used in the semester—as the fair score.[2] We define an assigned grade as harsh or lenient based on its difference from the fair score. By comparing the actual assigned grade to the fair score, the harshness or lenience of an individual grade is random by design. With this, we are able to analyze the causal effect of the harshness or lenience of a grade on the likelihood that the student submits a grade contest.

During the semester, when the students' scores were posted, we also included a generic female- or male-sounding name as their grader.[3] The name appears next to the student's grade and the scoring rubric. These names were randomly assigned and are not tied to any specific grader. After the scores and names were revealed, we asked the students if they thought they had been graded

1. This variation is not large on average. Appendix Figure B3 presents a histogram showing the distribution of deviations between grader-specific scores and the assignment average scores across 186 grader-semester-course-assignment combinations. Our focus here is not on extreme grading errors but rather on systematic differences among graders. To explore this further, we later test the robustness of our results using grader fixed effects. We find that when the same grader is associated with a female-sounding name reveals the same pattern.

2. One can think of this measure as the average (or expected) grade a student would receive from a random draw of possible grades.

3. These names were carefully selected. They are not the names of any faculty member or graduate student in the department. They are intentionally non-ethnic or tied to any specific culture. The names are Amanda, Emma, Eric, James, Jessica, John, Katherine, and Michael.

harshly or leniently.

This approach allows us to cleanly identify and measure the effect of being randomly assigned to a grader with a female- or male-sounding name, with varying degrees of harshness, on students' likelihood of contesting their grade and their opinion of the grade. Similar to a correspondence or audit study, randomly assigning female- or male-sounding grader names also allows us to isolate the effect of claimed gender of the evaluator from the effect of the inherently different feedback given by men and women[4].

We also explore whether the likelihood of contesting a grade is influenced by the potential differences in how male and female graders give feedback. Saygin and Knight (2023) find that female peer-graders tend to give lower scores than male peers in the context of peer evaluations, and Osun (2024) finds that women are more reluctant to give advice on difficult topics than men. Our analysis of the restricted sample in which randomly assigned female (male) names were attached to actually female (male) graders sheds some additional light on this question as well.

Our paper aims to shed light on the understanding of individual responses to feedback. It primarily contributes to a growing literature that explores gender disparities in performance evaluations. Specifically, we focus on the different reactions to feedback provided by female versus male evaluators. We find evidence that is consistent with prior research which demonstrates that holding performance constant, women in positions of leadership are evaluated more negatively than men, that is, they face more backlash (Blau and Kahn 2017; Boring 2017; Abel 2022; Born, Ranehill, and Sandberg 2022; Chakraborty and Serra 2024; Ayalew, Manian, and Sheth 2021; Elsesser and Lever 2011; Rudman et al. 2012; Reuben, Sapienza, and Zingales 2014;), are more heavily targeted on online review platforms (Rheault, Rayment, and Musulan 2019; Daniele, Dipoppa, and Pulejo

4. This study received IRB approval from University of Florida (IRB202201907). The data collection took place during the normal course of a college class. Displaying either a male or female name for the grader—unrelated to the grader's actual identity or gender—did not harm the students. The course instructors, to whom the students were required to submit grade appeals, were aware of the experiment, and the students had no interaction with their actual or named graders. Students were also clearly informed that all regrading decisions would be made by the course instructor. One might be concerned that the students could believe the instructor held a bias in favor of a certain gender of graders (e.g., "female graders always get writing scores correct, so their professor would deny any appeal that suggests a female made a mistake"), which might discourage the students from contesting errors. However, if the grader names were accurate, students would be just as likely—or unlikely—to be harmed by such a belief. Therefore, falsifying the names introduced no additional harm. If anything, by randomizing both the graders and their associated names, any potential harm would also be randomized. Finally, as the study was conducted within a single classroom context, it is unlikely to have contaminated subject pools commonly used in similar research. In addition to posing no harm, our results show that this minimal deception was necessary to detect the observed bias.

2023; Wu 2018), and are considered less credible and receive less recognition for their contributions and are punished more for their mistakes: (Dupas et al. 2021; Abel et al. 2024; Egan, Matvos, and Seru 2022; Sarsons 2022; Sarsons et al. 2021;Grossman et al. 2019; Carvalho 2025).

Previous work has also explored if female and male individuals differently seek out and respond to feedback. Coffman and Klinowski (2025) studied differences in *a priori* beliefs of performance and the demand for evaluative feedback. They observe that males have more optimistic *a priori* beliefs. However, there is no apparent gender gap in the demand for feedback. With respect to the perception of evaluative feedback, Roberts and Nolen-Hoeksema (1989) found that women perceive feedback, particularly negative feedback, to be more informative of their actual ability. Relatedly, they find that womens' self-assessments of their own abilities are influenced more by evaluative feedback than are mens' self-assessments. Shastry, Shurchkov, and Xia (2020) find that women are more likely to attribute negative feedback to an actual lack of ability, even when the feedback is due to bad luck. In other work, using surveys of randomized editorial decisions for submissions to top economics journals, Shastry and Shurchkov (2024) find that relative to an R&R, female assistant professors who receive a rejection perceive a significantly lower likelihood of subsequently publishing the paper in any leading journal than comparable male assistant professors. In an educational setting, male students are more likely to ask for and receive favorable regrades (Li and Zafar 2023) while in a professional setting, previous studies have also shown that women are less likely to negotiate: (Bowles, Babcock, and Lai 2007; Leibbrandt and List 2015, Small et al. 2007, Recalde and Vesterlund 2020; Roussille 2024). However, women know when to ask: they enter negotiations resulting in positive profits and avoid negotiations resulting in negative profits (Exley, Niederle, and Vesterlund 2020). Lastly, more recent papers explored the role of expected gender discrimination which may affect educational and professional decisions and outcomes (Alston 2019; Dustan, Koutout, and Leo 2022; Aksoy, Chadd, and Koh 2023; Ruebeck 2025; Gagnon, Bosmans, and Riedl 2024; Koutout 2022; Ugalde Araya 2024; Lepage, Li, and Zafar 2025; Exley et al. 2024). All of these studies provide compelling evidence that men and women may perceive harsh feedback differently, which may in turn affect how they respond.

We build on this growing literature by studying the reactions to feedback in an educational setting. Our experimental design, including both random assignment of multiple evaluators and

randomly assigned generic female- and male-sounding names for the graders, provides a unique opportunity to explore how the claimed gender of the evaluator impacts reactions to lenient or harsh feedback; that is, how it impacts resistance to this feedback by attempting to renegotiate the score by contesting the grade. It also offers several additional advantages. First, we separately observe students' reactions when they see a randomly assigned female-sounding name as their grader, and when they are actually evaluated by a female grader. We can isolate the differences in students' reactions that are driven by the unobserved differences in the feedback given by female versus male graders. Second, we collected a measure of students' confidence in their performance, which may affect their willingness to request regrading. Specifically, students provide grade predictions (expectations) when they submit each assignment. They are incentivized by an extra credit opportunity to provide accurate predictions. As a result, we identify how students' confidence in their work affects their reactions. Third, students were evaluated both based on the substance of their essays and their writing quality. We can separately examine students' responses to feedback on content (more objective) and writing quality (more subjective). Finally, we are able to explore the heterogeneity in these reactions based on several individual characteristics of the students and submitted essays.

Our results identify stronger resistance to feedback when the feedback is harsher than the fair score, and when students receive a lower grade than their expected score. Crucially, these complaints are more likely to occur when the evaluator is perceived as female. The stronger resistance to feedback provided by a female evaluator seem to be driven by both male and female students while being slightly stronger among male students. These results are similar for content and writing tasks implying that there is no apparent difference between the more objective, male-coded and the more subjective, female-coded evaluation items. However, we do find heterogeneity by the gender of the student: female students are more likely to contest female graders for writing scores while male students are more likely to contest female grader for content scores.

We also explore the role of ambiguity in SWA grades. While some assignments clearly indicate correct or incorrect answers, others involve partially correct responses that leave room for interpretation and variation in partial credit. To capture this ambiguity, we construct a measure based on the standard deviation of regrades from additional graders assessing the same submission. Using

grader disagreement as a proxy for ambiguity, we find that grade contests are more common for ambiguous submissions, and the gender bias against female-sounding grader names is concentrated in these cases—emerging in medium and high ambiguity assignments but not in clearly scored ones.

Finally, we replicated our analysis for our "*partial sample*." This is the restricted sample that includes only the (approximately) half of the observations in which the scores with a randomly assigned female-named grader is actually graded by a female, and the scores with a randomly assigned male-named grader is actually graded by a male. In this restricted sample, we find the same increase in grade contests for assignments that are evaluated by a female grader. However, this effect disappears once we control for the lenience or harshness of the scores, relative to the fair grade. These results suggest that we can only precisely identify the effect of evaluators' perceived gender on resistance to feedback by randomly assigning female- and male-sounding names to the graders.

We demonstrate the robustness of our main finding through a series of complementary analyses using alternative specifications and sample restrictions. These include excluding repeat contesters, contests missing the correct grader names, outlier grades, and regraded assignments, as well as modifying covariates and adding grader fixed effects. Across all specifications, the estimated effect remains stable. We also verify that the results are not driven by any particular name: jackknife analyses and conditional contest probabilities by name reveal no outliers, reinforcing that our findings reflect perceptions based on gendered name signals rather than individual identifiers.

Together, these results illustrate the gender biased resistance to harsh feedback in an educational context where we are able to isolate other confounding factors. While we are not able to provide direct evidence on the mechanisms, we are able to rule out a few alternatives. Since we eliminated student-grader interactions and required students to contact their instructor to contest a grade, the results cannot be driven by student beliefs that female graders would be more likely to accept a grade contest. While it is possible that the results are influenced by student beliefs that the course instructors (a male in both cases) would be more likely to approve a grade contest made against a female grader, our complementary findings suggest that students also *expect* female graders to be harsher. Students who contested their grade were disproportionately more likely to claim a grade was "harsh" *after* their grade contest was resolved, suggesting that students hold

their own beliefs, biased by the grader's gender, about a grade's validity.

The remainder of this paper is organized as follows. Section 2 describes our setting and experimental design. Section 3 outlines the empirical strategy, including the construction of the fair score. Section 4 presents the results. Section 5 provides a brief discussion and concludes.

2 Experimental Design

We study resistance to feedback in an educational setting using students enrolled in *Principles of Macroeconomics* and *Principles of Microeconomics* at a public flagship university. We received IRB approval and we pre-registered with the American Economic Association (AEARCTR-0010898). Below we describe our research design. Additional details of the design and data collection process are available in the appendix.

Each of the two courses was taught by a single designated instructor, both of whom were male. These very large introductory economics courses jointly enroll around $2,500$ students per semester. They are required for all business, economics, and journalism majors, however students from all undergraduate colleges and majors are represented. The overwhelming majority of students enrolled in these courses (around 97% in a given semester) are not economics majors. The two course instructors carefully aligned the structure and difficulty of their courses. They used the same number and types of assessments–low stakes online quizzes (10 percent), several high-stakes multiple choice exams (60 percent), and short writing assignments (30 percent)–which were scheduled at the same times in the semester and with the same structure. The instructors adopted a common course grading scheme.

Our focus is on the short writing assignments (SWAs). Both courses required students to complete two SWAs. Each SWA asked the student to answer a multi-part question with an objectively correct answer and was scored on a 100-point scale. The assignment prompts were posted in the Canvas learning management system. Students were required to compose their answers in essay form and construct a graph.[5] They submitted them in a single PDF document to Canvas. Before submitting their assignments, all students were required to respond to a two-

5. To deal with the widespread availability of generative AI applications (e.g., ChatGPT), each prompt also required a hand-drawn graph, which then-available AI technologies could not handle well. Only the graph is handwritten and the essay is typed.

question survey asking for (1) acknowledgment that violation of certain submission requirements will result in a 0 on the assignment and (2) a grade expectation.[6] We incentivized students to honestly report their grade expectations by offering extra credit for accuracy.[7] The distribution of the deviations of students' expected grades from their actual grades and the fair grade measures are presented in the Appendix Figure B2. These deviations measure the extent to which students are surprised by their grades, as well as their prediction accuracy. There is no statistically significant difference between graders with male-sounding names and graders with female-sounding names in the deviation between expected scores and fair scores. The difference between the expected score and the fair score is statistically insignificant, though on average students overestimate their scores and this overestimation is higher among male students. This finding is consistent with the previous studies that identify more pronounced overconfidence among male students (Niederle and Vesterlund 2007, Azmat and Petrongolo 2014, Exley and Kessler 2022, Bordalo et al. 2019, Buser, Niederle, and Oosterbeek 2014, Coffman 2014, Baldiga and Coffman 2018, Iriberri and Rey-Biel 2021, Saygin and Atwater 2021).

Each grader was given a random set of assignments after the submission deadline and was instructed in what order to evaluate the assignments. This order was also randomized, and we monitored the graders to ensure that they evaluated submissions in the assigned order. Graders were provided with a clear scoring rubric. Each rubric includes ten items, each of which is worth ten points. There are seven items related to the economics content and three items related to the essay's writing quality. The assignment prompts are written to solicit answers to seven specific content questions to which there are objectively correct answers. For example, one prompt presented a linear production possibilities frontier between blueberries and raspberries and asked the students to calculate the opportunity cost of producing an additional unit of blueberries. There is a single correct answer to this content question. However, graders were allowed to assign different amounts of partial credit in some circumstances, such as when a student presented the correct math but an incorrect final answer or when a student reported a correct answer using incorrect units. The

6. The acknowledgment includes affirmations that their name is NOT included in the assignment document or file name and that their work is typed and formatted as a PDF.

7. Students were provided extra credit on the SWA if their guess was within 5 points of the TA-assigned grade and an additional point if their prediction exactly matched the TA-assigned grade. A screenshot of the exact language provided to students is included in Appendix Figure A3.

graders also evaluated the answers' writing quality. The three writing quality items in the rubric are open to a more subjective evaluation. For example, the grader is asked to rate the overall writing quality of the essay on a ten-point scale. As with the content questions, the graders could assign partial credit for the writing quality questions.

Grading is done blind. Students are told not to include their name on their assignment submission and were penalized if they violated this instruction. In the rare event that a student included their name, we flagged the submission and excluded it from the analysis. Canvas settings also hide students' names and generate an otherwise meaningless student number for each submission. Only after the grades were finalized and released to the student could the instructors and graders identify the author of each submission. Additionally, at no point did the graders interact with the students. They did not hold office hours and they did not respond to student inquiries (e.g., grade challenges).

Once all the grading was complete, the scores were released to the students. At this time, the graders can no longer revise the scores. The students observe the number of points that they received on each of the ten items in the scoring rubric and also see a randomly assigned name of a grader. These names are not the names of the actual graders, nor are they the names of any faculty member or graduate student in the department. These names include four female- and four male-sounding names: Amanda, Emma, Eric, James, Jessica, John, Katherine, or Michael.

If a student wished to contest their grade, they were required to submit the grade contest to their instructor.[8] Grade contests were required to be submitted within one week after assignment scores were released. Students had to also include the (randomly assigned) claimed name of the grader in their grade contest to the instructor, which ensured that the student saw the "name" of their grader. All grade contests were archived. Data on these contests, including any grading errors and score corrections, were collected by the instructor and analyzed after the semester concluded.

In Appendix Figure A1 we provide a weekly timeline of the experiment over the typical semester. But in brief, the SWAs were open to students for one week. The grading process began within 24 hours after the submission deadline. The grades were released to students one week later, shortly after the last grader provided their grades to the instructor. Once the scores were released, students

8. They had no way to contact their grader as the actual graders were not included in the elearning platform and their contact information was undiscoverable.

had one week to submit grade contests. Finally, when the students completed their next SWA, they were asked to rate the overall fairness of the grade they received on the *previous* assignment. The exact text of this question is provided in Appendix Figure A2.

Because we want to examine how the lenience or harshness of the grades affect students' tendency to contest their grades, we need a measure of a "fair" grade for each SWA submission. We obtained this fair grade by having additional graders evaluate each submission, *ex-post*, after the semester ends. These additional graders received the same training and guideline as the initial graders who assigned the students' scores during the semester. We use the average of these three grades, the two new *ex-post* grades and the grade assigned during the semester to approximate a fair grade. The harshness or lenience of the actual assigned grade is measured as the deviation from this fair grade.[9] This allows us to test whether and how students' grade contests vary depending on the harshness of the grades they receive, as well as whether there is any heterogeneity in these reactions based on the claimed gender of the grader.

3 Sample Selection and Empirical Approach

We examine whether students react to their grades differently based on the perceived gender of their evaluator. Moreover, we explore whether being randomly assigned to a tougher grader influences students' reactions to their grades, and whether these reactions also depend on the perceived gender of the grader.

It is difficult to identify whether individuals find performance evaluations to be unfair and/or challenge this feedback precisely because they are evaluated by a female supervisor. One common confounding issue is presented by the sorting of individuals and their supervisors based on observed and unobserved characteristics, including the gender and "toughness" of the supervisor. We remove this concern by randomly assigning the actual graders, as well as the female- and male-sounding names we claim were the graders. Moreover, our experimental setting rules out the possibility that graders' evaluations or students' contest decisions are affected by grader-student interactions. Our graders do not interact with the students at any point during the course.[10]

9. The distribution of this measure, separated by the claimed gender of the grader, is presented in the Appendix Figure B1.

10. While the role that supervisor-supervisee interactions play in performance evaluations and reactions to feedback

Our experimental design separates students' reactions that are attributable to the claimed gender of the grader and students' reactions that are attributable to the differences in feedback that is given by female and male evaluators. In our setting a gender biased reaction would be characterized by a higher share of grade contests arising when the experimental name of the grader is female-sounding, regardless of the actual grader's gender. If, on the other hand, reactions to feedback are driven less by the grader's gender *per se* and more by differences in unobserved characteristics of female graders to provide feedback, we would expect that the grade contest differences would be similar by actual graders' gender (which is not observed by students).

We employ an empirical framework that predicts the probability that a student contests their grade based on the claimed gender of the grader, which is based on the randomly assigned female- or male-sounding name. We also explore how the fairness of the grade impacts these grade contests. We use the measure for a fair grade that was described in the previous section and characterize the "lenience" (or "harshness" if negative) of a grade by its deviation from this fair grade[11]:

$$Lenience_{ijcs} = ActualGrade_{ijcs} - \underbrace{\frac{1}{n_r+1}\left(ActualGrade_{ijcs} + \sum_{r=1}^{n_r} Regrade_{ijcs}^{(r)}\right)}_{FairGrade_{ijcs}}$$

where $Regrade_{ijcs}^{(r)}$ denotes the rth *ex-post* score given to the submission provided by student i on assignment j in semester s of course c. In our case we utilize two *ex-post* regrades, such that $n_r = 2$. While it may seem intuitive to exclude the actual score from the fair grade calculation, both methods are unbiased and including all available scores is more efficient.[12]

In our empirical specification, we predict the probability of exhibiting resistance to feedback by submitting a grade contest for an assignment submitted by student i for SWA j in semester s of course c:

$$Contest_{ijsc} = \alpha + \beta ClaimedFemale_{ijsc} + \mathbf{G}_{ijsc} + \mathbf{X}_{ijsc} + \nu_j + \mu_s + \eta_c + \epsilon_{ijsc} \quad (1)$$

where $ClaimedFemale_{ijsc}$ is an indicator for whether student i's grade on assignment j in semester

is important in better understanding the nature of the responses to feedback, it is beyond the scope of this paper.

11. The distribution of this measure by the claimed gender of the grader is presented in the Appendix Figure B1.

12. This is further shown in Appendix B.

s of course c was randomly assigned a female-sounding grader name. $\boldsymbol{G}_{ijsc}$ is a vector of grade-specific controls. In our preferred specification this includes the lenience of a grade as defined above, separately for writing and content sub-scores. In other specifications, we use the actual score, fair grade, grader fixed effects, and actual grader gender. $\boldsymbol{X}_{ijsc}$ is a vector of student-specific controls, including student gender and expected scores. Our baseline specification also includes assignment, semester, and course fixed effects ν_j, μ_s, and η_c, respectively. β measures biased reactions to grades that are driven by the graders' claimed gender.

Our analysis is conducted on the sample of student essays that are correctly submitted following the assignment instructions. There were 4133 unique students over 2 SWAs, which generated 8523 unique assignment submissions.[13] We exclude 25 submissions, or less than 0.3% of all submissions, due to either a grading or submission error and make no other restrictions.[14]

4 Results

We start our analysis with a confirmation of randomization of the grader's names across SWAs submitted by students. Table 1 shows balance on several variables by claimed gender of the graders. We have only one variable that shows a significant difference. The share of assignments completed by female students is higher among the assignments claimed to be graded by a male-sounding name. In all of our baseline specifications, we control for the gender of the student to take this difference into account. Of the 8498 graded submissions, 439 (or 5.17%) were contested to the instructor. Of these 439 contests, 206 (or 46.92%) mentioned the writing score as part of their complaint, while 387 (or 88.15%) mentioned content-specific elements of the rubric.

Figure 1 presents the share of grade contests overall and by male and female students by the claimed gender of their grader inferred from the female or male-sounding names. Panel A presents all grade contests, while Panel B and C show the share of grade contests for the content and writing scores respectively.

These graphs reported in Figure 1 reveal that students are more likely to contest their grade

13. Some students did not submit a particular assignment or provided a file which could not be opened. This results in 59 dropped observations.

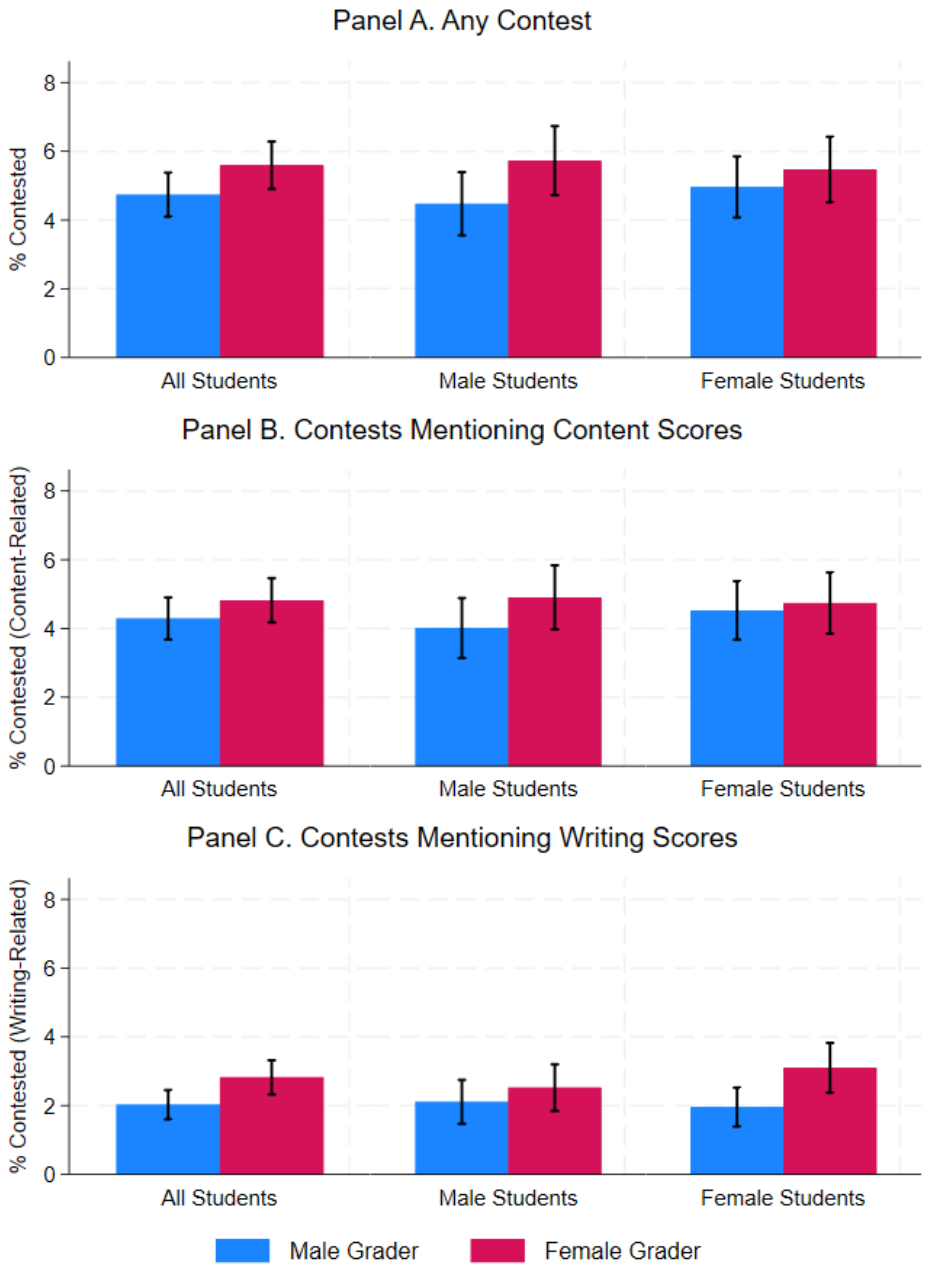
14. We exclude 11 submissions that did not follow the assignment instructions on submitting a typed essay, 12 submissions that were either submitted late or included the student's name and were not graded blindly, and 2 submissions whose grade exceeded the maximum possible score due to a grading error.

Table 1: Summary Statistics with Sample Restrictions by Claimed Grader Gender

	(1)	(2)	(3)
Variable	Male Claimed Grader	Female Claimed Grader	Difference (se)
Female Student	0.542	0.516	-0.026**
	(0.498)	(0.500)	(0.011)
Female Actual Grader	0.503	0.516	0.013
	(0.500)	(0.500)	(0.011)
Actual SWA Score	92.394	92.171	-0.223
	(10.439)	(10.738)	(0.230)
Fair Grade	92.411	92.136	-0.275
	(9.625)	(10.157)	(0.215)
Expected SWA Score	93.795	93.710	-0.084
	(5.919)	(6.152)	(0.131)
Actual - Fair Grade	-0.017	0.035	0.052
	(3.985)	(4.051)	(0.087)
Actual - Expected Grade	-1.401	-1.539	-0.138
	(10.377)	(10.789)	(0.230)
Fair - Expected Grade	-1.384	-1.574	-0.190
	(9.606)	(10.095)	(0.214)
SD of Regrades (Ambiguity)	2.025	2.030	0.005
	(3.101)	(3.191)	(0.068)
Contested Prior Grade	0.035	0.033	-0.002
	(0.183)	(0.179)	(0.004)
Course Grade after SWA2	80.996	80.874	-0.122
	(13.434)	(13.192)	(0.290)
Observations	4,244	4,254	8,498

Notes: Unsubmitted and improperly submitted assignments (those receiving a 0 on writing or those regraded due to name appearing) and clear grader errors are omitted.

Figure 1: Main Result by Claimed Gender



Note: This figure presents the share of grade contests of male and female students by the claimed gender of their graders defined by female or male-sounding names. Panel A shows overall contests, while Panel B shows contests on content scores and Panel C shows contests on writing scores. Unsubmitted and improperly submitted assignments (those receiving a 0 on writing or those regraded due to name appearing) and clear grader errors are omitted.

when their SWA is graded by a grader with a female-sounding name. This is true independent of whether the contest was made for the content or writing score as shown in Panels B and C. Interestingly, we observe that this gender bias in contests seem to be stronger among female students for contests in writing scores while it is stronger among male students for contests in content scores. This finding suggests that both female and male students' contests are biased against female graders, particularly in tasks where the students are arguably more confident. This is consistent with prior literature showing that women (men) tend to be more overconfident in tasks that are typically female- (male-) dominated. (For example, Bordalo et al. 2019 and Coffman 2014).

Next, we present the results from the main specification described in the previous section in Table 2. This table formalizes the patterns in Figure 1. We predict the probability of a grade contest from the inferred gender of the grader from the randomly assigned grader names. We include semester, course, and assignment fixed effects in Columns 2 and 5. Finally, we add further controls for the leniency of content and writing scores they receive as well as the gender of the student. In Columns 1-3 we report the results from the first SWA and in columns 4-6, we pool SWA1 and SWA2 together. While the point estimates in the two samples are very similar, we obtain more precision with the increased sample size.[15] These results suggests that graders with female-sounding names face more resistance to the feedback they provide regardless of their actual gender and even conditional on the leniency of scores they assign. Specifically, these assignments graded by female-labeled graders brought a 0.96 percentage point, or approximately 20%, higher chance of being contested to the course instructor.

In order to provide evidence for the robustness of this finding, we conduct a series of additional analysis. In Figure 2 we report the point estimates of the effect of reporting a female-sounding name as the grader on the probability of grade contest together with the confidence intervals under different specifications and sample restrictions. We start by repeating columns (4)-(6) of Table 2: the bivariate analysis which provides us with unconditional mean differences (column 4); course, assignment, and semester fixed effects (column 5); and additional controls for writing and content lenience and student gender (column 6). We refer to the latter as our "Baseline" specification. In

15. Given that students do no often contest their grades, we have limited variation in our outcome variables and this leaves us with limited power to identify the potential differences precisely. However, increased sample size improves the precision once we analyze the two SWAs together.

Table 2: Linear Probability of Contesting Grade to Instructor

	SWA 1 Only			SWA 1 & 2		
	(1)	(2)	(3)	(4)	(5)	(6)
Female Claimed Grader	0.923	0.975	1.084	0.859*	0.866*	0.962**
	(0.772)	(0.766)	(0.731)	(0.480)	(0.477)	(0.463)
Writing Lenience			-1.040***			-1.210***
			(0.338)			(0.207)
Content Lenience			-1.994***			-1.390***
			(0.211)			(0.128)
Female Student			-0.255			0.157
			(0.733)			(0.464)
Fixed Effects	No	Yes	Yes	No	Yes	Yes
Mean Dependent Variable	6.821	6.821	6.821	5.166	5.166	5.166
Observations	4266	4266	4266	8498	8498	8498

The dependent variable takes value 100 if students contested the grade to their instructor and 0 otherwise. Fixed effects include separate course, assignment, and semester-specific controls. Robust standard errors are presented in parantheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

the rows which follow, we test this baseline specification under various sample restrictions. First, we exclude all SWA 2 submissions of those who previously contested SWA 1 in the same course-semester. Those students who contested the first SWA and were denied may be less likely to contest again on the second assignment, biasing downward our results. Alternatively, some students may be inclined to always contest their grade, and, if these students were more likely to receive a female-sounding grader, this may bias upwards our result. Given this sample restriction, the result is little changed. Next, we exclude all those submissions that did not include the correct grader name in their contest email (if they contested). This is an aggressive way to reduce potential noise stemming from students not actually observing the grader name before contesting. Indeed, the confidence interval narrows, and the point estimate is little changed. We do not make this exclusion in the baseline sample as some students may be aware of the grader but were either unaware of the requirement to include the grader's name, simply forgot to include it, or strategically excluded it. We also exclude outlier submissions which received scores far from the fair grade either in content scores or overall and particularly low-performing submissions (fair grades below 70). The former also accounts for potential grading errors which would generate valid grade contests. In all three cases, the point estimate is little changed.

Figure 2: Robustness Analysis for the Linear Probability of Contesting Grade to Instructor with Various Sample restrictions and Specifications

Note: This figure presents the robustness analysis of the baseline estimates. "Bivariate", "Bivariate + FE", and "Baseline" refer to columns (4)-(6) of Table 2, respectively. In the subsequent rows, we exclude SWA 2 submissions of those who contested on SWA 1, all submissions of those who did not include the correct name of the grader in their contest email, submissions that received scores more than 20 points away from the fair content or fair overall grade, all particularly low performing submissions (fair grades below 70), contested submissions for which the grade contest was accepted, or those receiving a score ending in an 8 or 9 (near a cuttoff/ left digit bias). In the final two rows, we replace the lenience controls in writing and content with actual writing and content subscores and grader-semester fixed effects. Whiskers present 95% confidence intervals using robust standard errors.

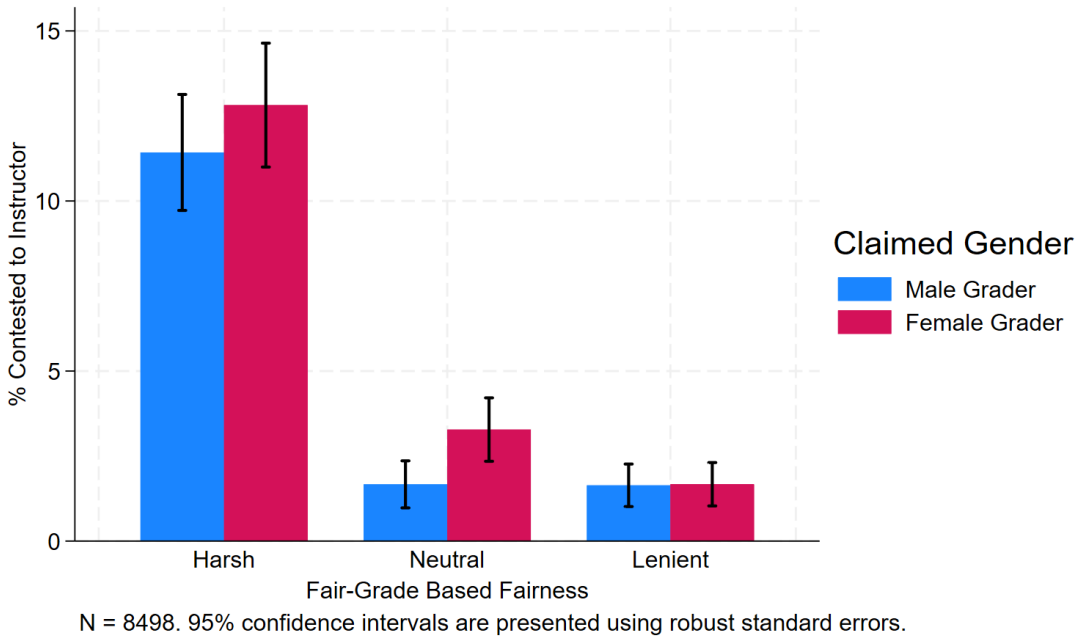
As stated above, some grade contests are valid complaints due to genuine grading errors. Maintaining these in the data add noise to our estimates, so we also include a specification which removes all regraded submissions. Our estimates are more precise and virtually identical to the baseline sample. We do not make this restriction in our baseline sample as there may be valid concerns that course instructors, who are aware of the experiment, may be subconciously inclined to systematically approve certain contests.

If a student's decision to contest a score is a function of the perceived award from a successful contest, then one may expect students to contest more near certain score cutoffs. If the submissions that score near these cutoffs were more/less likely to be assigned a female-sounding name, this would bias our results. In the third row from the bottom, we exclude all submissions whose score ends in an "8" or "9." The restriction reduces our sample size and power significantly, but the coefficient estimate is again little changed. Lastly, we alter our empirical specification to use actual writing and content subscores (rather than the respective lenience measures) and, in the final row, add grader fixed effects to account for grader-induced variation in scores. In either specification the results are little changed.

As an additional robustness check we verify that our results are not driven by any particular name, which would suggest that our interpretations are misguided. In Appendix Figure C1 we present the results of two checks. First, in subfigure (a), we plot the conditional probabilities that each name is contested and show that the male and female names cluster around similar probabilities with no clear outliers. In subfigure (b) we plot each of 8 jacknife coefficients produced from separate estimations of our main regression specification, each omitting all observations which were assigned a particular grader name. None of the 8 estimates are statistically different from our main specification, lending support to our interpretation of the results as between male-sounding and female-sounding names and not any other common signal.

Next, we elaborate on the role of leniency or harshness of grades on the contesting behavior. We question whether increased grade contests for female-sounding name graders depends on the leniency of grades assigned by these graders. We define lenience based on the difference between the actual score they receive and the 'fair grades' proxy calculated as the average of the 3 scores assigned to the same SWA by 3 different graders. A score is considered "neutral" if it is within 1

Figure 3: Grade Contests by Lenience and Claimed Gender



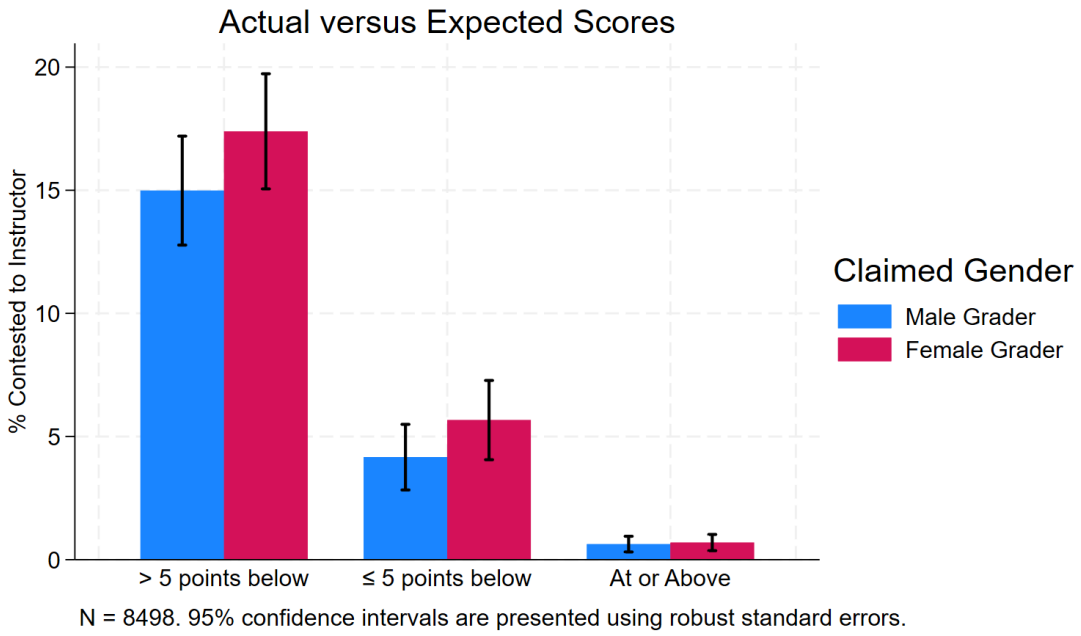
Note: This figure presents the effect of being graded by a grader with a female-sounding name on different values of grade lenience measured as the difference between the actual score they receive and the 'fair grades' proxy calculated as the average of the 3 scores assigned to the same SWA by 3 different graders. A score is considered "neutral" if it is within 1 point of the "fair" grade.

point of the "fair" grade. Any grade above (below) that would be lenient (harsh). Figure 3 shows that most of the grade contests are among the SWAs that were graded harsher than the average and the bias against graders with a female-sounding name is mostly driven by the groups of SWAs that were graded harshly or neutrally while we observe no difference among those that were graded leniently.

To dig deeper, we also analyze the deviation of the assigned grades from the students' expected grades. We test whether students are more or less likely to contest their grades when they receive a lower or higher score than what they expected.

As described in the previous sections, we elicited the expected scores with incentives when they submitted their SWAs before the grading starts. In our next analysis, we focus on the difference

Figure 4: Gender Bias in Grade Contests by Grade Shocks

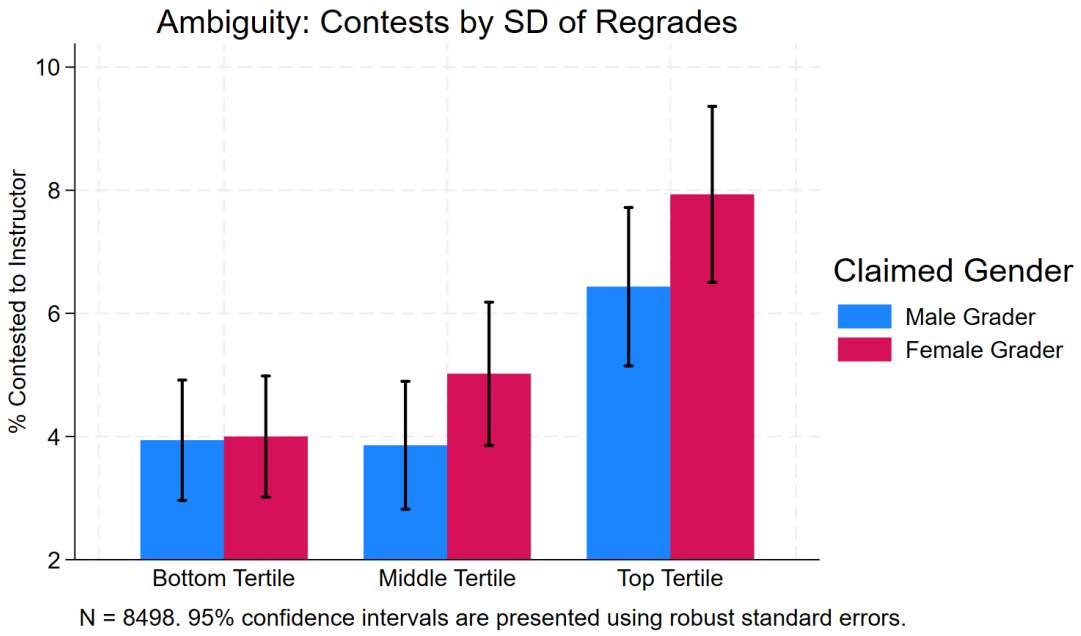


Note: This figure presents the effect of being graded by a grader with a female-sounding name on those receiving grades more than 5 points below, less than 5 points below, or at/above their expected score.

between the actual scores they receive from the expected scores and we test whether the effect of being graded by a grader with a female sounding name varies for different values of this difference. Figure 4 reveals that the gender bias in grade contests is mostly driven by students who received lower scores than they expected and the estimates become insignificant when they receive higher scores than expected.

Another dimension we examine in our data is the ambiguity of SWA grades. Some SWAs offer little room for partial credit, as they clearly indicate a correct or incorrect answer. Others, however, contain responses that are partially correct, leading to differing interpretations and variations in the amount of partial credit assigned by graders. To capture this, we constructed a measure of ambiguity based on the disagreement among additional graders evaluating the same SWA, quantified by the standard deviation of regrades. We then analyzed the role of this ambiguity in grade contest behavior, reporting results separately for three groups—bottom, middle, and top

Figure 5: Gender Bias in Grade Contests by Ambiguity of Grades (SD of Regrades



Note: We constructed a measure of ambiguity using the standard deviation of regrades assigned by additional graders to the same SWA, capturing the level of disagreement among graders on the same homework submission. This graph displays the analysis of grade contests separately for three groups—bottom, middle, and top tertiles—based on the distribution of this standard deviation.

tertiles—based on the distribution of standard deviations. The top tertile reflects SWAs with the highest grader disagreement. Figure 5 shows that grade contests are more frequent for SWAs with higher ambiguity (top tertile). Moreover, gender bias against female-sounding grader names begins to emerge in the middle and top tertiles, while no significant difference is observed in the bottom tertile.

Finally, we explore the role of potential gender differences in grading style by analyzing the subsample of SWAs that were claimed to be graded by a female-sounding name and indeed were graded by a female grader. As the graders are randomly assigned to the SWAs and the female and male-sounding names are randomly assigned to graded SWAs, approximately half of the SWAs were graded by a female grader and also had a female-sounding name visible to the student as grader. We replicate our results for this partial sample with true and false genders in Table 3.

Table 3: Regression Results for Different Samples: Linear Probability of Contesting Grade to Instructor

	A. Full Sample				B. Partial Sample			
	Stated Gender (1)	Stated Gender (2)	Actual Gender (3)	Actual Gender (4)	True Gender (1)	True Gender (2)	False Gender (3)	False Gender (4)
Female Claimed Grader	0.859*	0.962**			0.844	0.509	0.914	1.322*
	(0.480)	(0.463)			(0.641)	(0.611)	(0.717)	(0.692)
Writing Lenience		-1.210***		-1.198***		-0.856***		-1.603***
		(0.207)		(0.207)		(0.287)		(0.275)
Content Lenience		-1.390***		-1.393***		-1.388***		-1.404***
		(0.128)		(0.128)		(0.184)		(0.176)
Female Student		0.157		0.134		0.275		0.014
		(0.464)		(0.464)		(0.620)		(0.693)
Female Actual Grader			-0.025	-0.382				
			(0.480)	(0.461)				
Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Observations	8498	8498	8498	8498	4303	4303	4195	4195

The dependent variable takes value 100 if students contested the grade to their instructor and 0 otherwise. The mean of the dependent variable is 5.17. Fixed effects include separate course, assignment, and semester-specific controls. Robust standard errors are presented in parantheses. * p < 0.10, ** p < 0.05, *** p < 0.01

Columns 1 and 2 of Table 3 Panel A report our main findings using the full sample, included here for comparison purposes. Columns 3 and 4 use the same specification as our main analysis but replace the claimed gender with the gender of the actual grader, again using the full sample. We find no evidence of an increased probability of grade contests for the SWAs that were actually graded by a female grader. In fact, the point estimates are negative, insignificant, and sensitive to the inclusion of additional controls for the leniency of the assigned score.

Panel B of Table 3 replicates our main analysis for the "partial sample." The first two columns of the panel focus on "true gender," in which we restrict the sample to SWAs that were both graded by a female (male) grader and also had a female-sounding (male-sounding) name visible to the student as the grader. The results for this partial sample are similar to those found in the full sample (column 1 of Panel A). While the precision of the point estimates are lowered, we still find that students are more likely to contest their grade when they are graded by a female grader reported with a female-sounding name as grader. However, once we control for the leniency

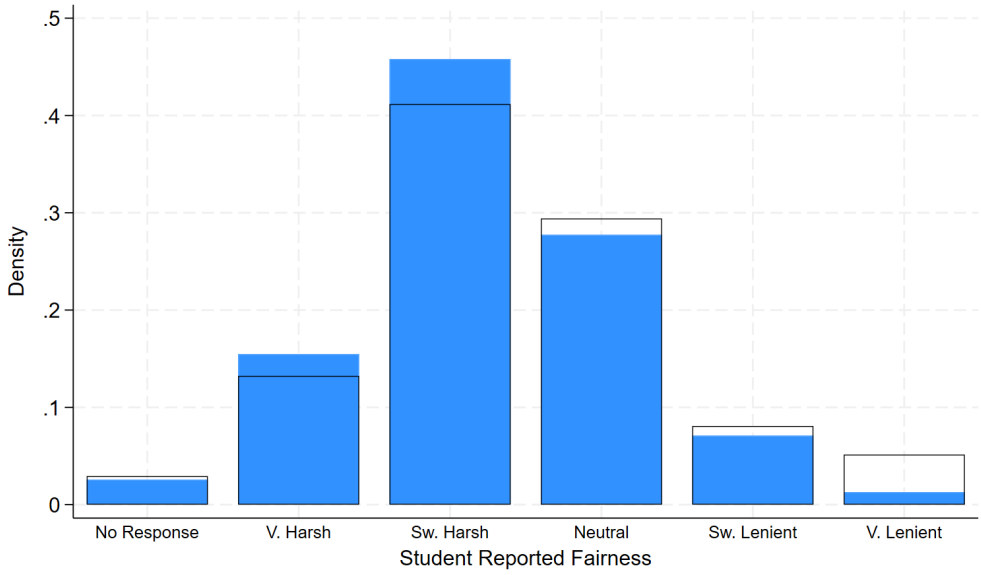
of the assigned grades in the next column, this gender difference becomes substantially smaller and statistically insignificant. This sensitivity to leniency controls suggests that our female and male graders differ in their grading leniency in ways that directly impact students' likelihood of challenging their grades.

Finally, the last two columns repeat the same analysis on the sample where students observed a female-sounding (male-sounding) name, but the actual grader was of the opposite gender. Without leniency controls, we again find very similar results to our main specification. However, once we control for grading leniency, the point estimates increase relative to column 2. This implies that differences in grading style or standards of female and male graders do play some role in the observed patterns in grade contests. Importantly, our main specification, which uses randomly assigned female- and male-sounding names, helps rule out these grader-level effects.

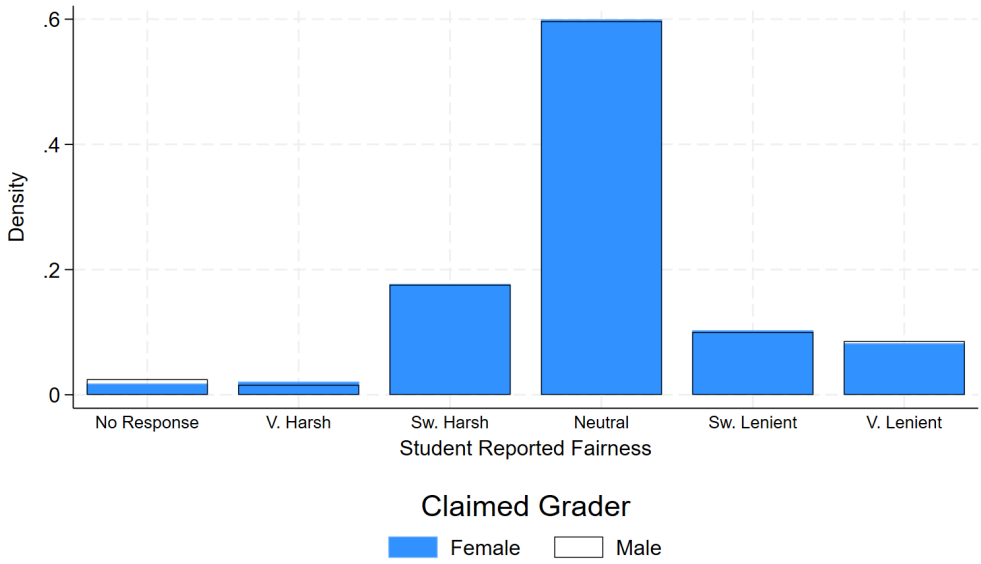
The bias we identify here may be due to several factors which we will now explore. First, we can rule out student beliefs that female graders would be more likely to accept a grade challenge as we require all grade challenges to be submitted to the course instructor. Alternative explanations for this behavior are that students may: (1) believe male instructors are more likely to approve a grade contest against a female grader or (2) that students believe the grades provided by female evaluators are less valid, either by being harsher or more prone to mistakes. While we are not able to completely rule out either explanation, we do find evidence in support of (2). When students went to submit their second SWA, they were asked to rate the fairness of the grade they received on the first SWA, which had been graded nearly a month prior and for which all grade contests had been resolved. The exact language of this question is provided in Appendix Figure A2. The results of this survey are presented in Figure 6. Students who contested their score were more likely to report the grade they received as "somewhat harsh" or "very harsh." Among those who did not contest (Panel b), there is no difference in how students reported the grades provided by graders we claimed were male or female. However, among those who did contest (Panel a), there is a large shift away from "lenient" and toward "harsh" for female graders. We interpret this difference to mean that students hold beliefs about the validity of a grade, and they hold this belief even after any grade contests have been resolved.

Figure 6: Distribution of Fairness Question Responses

(a) Conditional on Contesting Score



(b) Conditional on Not Contesting Score



Notes. Both figures plot the distribution of student responses to a question about the grading of SWA 1 which was asked when students submitted SWA 2. Figure (a) is restricted only to those who contested their SWA 1 score while Figure (b) is restricted to those who did not contest. Unsubmitted and improperly submitted SWA 1 assignments (those receiving a 0 on writing or those regraded due to name appearing) and clear grader errors are omitted.

5 Discussion

Individuals perceive and respond to feedback on their performance differently. These responses may affect their future outcomes in educational and professional settings. Our paper identifies bias in students' reactions to grades in a unique experimental setting. We find that students are more likely to contest a grade provided by a female-sounding evaluator than a similar grade provided by a male-sounding evaluator. This disparity is most pronounced when grades are harsh relative to a "fair" assessment, are below students' expectations, or when the submission produces ambiguity in grading.

We conduct our experiment in two large introductory economics courses at a public flagship university. Students submit two short writing assignments in each course. These assignments are graded blindly by randomly assigned graders. Separately, students' grades are displayed with a randomly assigned generic name, which may be a female- or male-sounding name. This setting allows us to isolate the effect of the grader's perceived gender from other observed and unobserved characteristics, which may be correlated with how female and male graders differently evaluate assignments. We show that this separation is essential for identifying and precisely measuring bias in students' resistance to feedback, because the graders' actual gender correlates with other observed characteristics.

Our study lays the groundwork for future research into the reasons for resistance to feedback that is believed to have been provided by a female evaluator. In this paper, we find that students are more likely to contest grades that are provided by women, particularly when they are graded harshly, and when they receive a lower score than they expected, and when there is more disagreement among the additional graders of the same assignment. These observations are consistent with a mechanism in which students may tend to believe that female graders are more likely to make mistakes. They are also consistent with a mechanism in which students have the belief that female graders are more likely to give unreasonably harsh grades. In both instances, students would be expected to contest grades provided by female graders to their instructor more often.

It is reasonable to expect that the observed differences in grade contest rates by the claimed gender of the grader are driven by students' expectations of success in securing a favorable regrade. However, our experimental design largely rules out this mechanism—where students may be more

likely to contest grades assigned by female graders because they are perceived as more lenient or more likely to grant regrade requests. In our setting, all regrade requests are submitted directly to the course instructors, who evaluate and decide on them independently.

That said, we cannot rule out the possibility that students believe their instructors may perceive female graders as more error-prone or harsher. Students' own biases and their expectations of similar biases in instructors may reinforce one another. In other words, students may believe that women are more likely to make grading mistakes, or they may think instructors are more likely to accept that a mistake was made when the grader is female.

While we do not directly test these mechanisms, our complementary analysis shows that students who contested their grade were disproportionately more likely to describe it as "harsh" *after* the contest was resolved. This suggests that students hold their own beliefs about a grade's validity which persists even after a grade contest is resolved.

Our findings may have important implications for understanding gender dynamics and biases in performance evaluations beyond educational settings. Stronger resistance to feedback from female evaluators, coupled with the role of grading harshness, suggests that perceptions of authority and fairness may be influenced by gender stereotypes. Addressing these biases could involve training supervisors and their employees to recognize and mitigate such tendencies. It could also involve facilitating greater transparency in evaluation processes. These steps would foster more equitable environments in professional and educational settings. They would also reduce the undue burden on female supervisors and educators.

References

Abel, Martin. 2022. "Do Workers Discriminate against Female Bosses?" *Journal of Human Resources* (April).

Abel, Martin, Emma Bomfim, Izzy Cisneros, Jackson Coyle, Song Eraou, Martha Gebeyehu, Gerardo Hernandez, et al. 2024. "Are women blamed more for giving incorrect financial advice?" *Journal of Economic Behavior & Organization* 228 (December): 106781.

Aksoy, Billur, Ian Chadd, and Boon Han Koh. 2023. "Sexual identity, gender, and anticipated discrimination in prosocial behavior." *European Economic Review* 154 (May): 104427.

Alston, Mackenzie. 2019. "The (perceived) cost of being female: An experimental investigation of strategic responses to discrimination." *Department of Economics, Florida State University, Tallahassee, FL. Accessed November* 21:2019.

Ayalew, Shibiru, Shanthi Manian, and Ketki Sheth. 2021. "Discrimination from below: Experimental evidence from Ethiopia." *Journal of Development Economics* 151 (June): 102653.

Azmat, Ghazala, and Barbara Petrongolo. 2014. "Gender and the labor market: What have we learned from field and lab experiments?" *Labour Economics,* Special Section articles on "What determined the dynamics of labour economics research in the past 25 years? edited by Joop Hartog and and European Association of Labour Economists 25th Annual Conference, Turin, Italy, 19-21 September 2013 Edited by Michele Pellizzari, 30 (October): 32–40.

Baldiga, Nancy R., and Katherine B. Coffman. 2018. "Laboratory Evidence on the Effects of Sponsorship on the Competitive Preferences of Men and Women." *Management Science* 64, no. 2 (February): 888–901.

Blau, Francine D., and Lawrence M. Kahn. 2017. "The Gender Wage Gap: Extent, Trends, and Explanations." *Journal of Economic Literature* 55, no. 3 (September): 789–865.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2019. "Beliefs about Gender." *American Economic Review* 109, no. 3 (March): 739–773.

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

Born, Andreas, Eva Ranehill, and Anna Sandberg. 2022. "Gender and Willingness to Lead: Does the Gender Composition of Teams Matter?" *The Review of Economics and Statistics* 104, no. 2 (March): 259–275.

Bowles, Hannah Riley, Linda Babcock, and Lei Lai. 2007. "Social incentives for gender differences in the propensity to initiate negotiations: Sometimes it does hurt to ask." *Organizational Behavior and Human Decision Processes* 103, no. 1 (May): 84–103.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek. 2014. "Gender, Competitiveness, and Career Choices *." *The Quarterly Journal of Economics* 129, no. 3 (August): 1409–1447.

Carvalho, Marcela. 2025. *Who gets the benefit of the doubt? CEO gender and news about firm performance.* Working Paper, June.

Chakraborty, Priyanka, and Danila Serra. 2024. "Gender and Leadership in Organisations: the Threat of Backlash." *The Economic Journal* 134, no. 660 (May): 1401–1430.

Coffman, Katherine, and David Klinowski. 2025. "Gender and Preferences for Performance Feedback." *Management Science* 71, no. 4 (April): 3497–3516.

Coffman, Katherine Baldiga. 2014. "Evidence on Self-Stereotyping and the Contribution of Ideas *." *The Quarterly Journal of Economics* 129, no. 4 (November): 1625–1660.

Daniele, Gianmarco, Gemma Dipoppa, and Massimo Pulejo. 2023. *Violence against Women in Politics.* SSRN Scholarly Paper. Rochester, NY, August.

Dupas, Pascaline, Alicia Sasser Modestino, Muriel Niederle, Justin Wolfers, and The Seminar Dynamics Collective. 2021. *Gender and the Dynamics of Economics Seminars.* Working Paper, February.

Dustan, Andrew, Kristine Koutout, and Greg Leo. 2022. "Second-order beliefs and gender." *Journal of Economic Behavior & Organization* 200 (August): 752–781.

Egan, Mark, Gregor Matvos, and Amit Seru. 2022. "When Harry Fired Sally: The Double Standard in Punishing Misconduct." *Journal of Political Economy* 130, no. 5 (May): 1184–1248.

Elsesser, Kim M., and Janet Lever. 2011. "Does gender bias against female leaders persist? Quantitative and qualitative data from a large-scale survey." *Human Relations* 64 (12): 1555–1578.

Exley, Christine L, and Judd B Kessler. 2022. "The Gender Gap in Self-Promotion*." *The Quarterly Journal of Economics* 137, no. 3 (August): 1345–1381.

Exley, Christine L., Raymond Fisman, Judd B. Kessler, Louis-Pierre Lepage, Xiaomeng Li, Corinne Low, Xiaoyue Shan, Mattie Toma, and Basit Zafar. 2024. *The Gender Concealment Gap.* Working Paper, April.

Exley, Christine L., Muriel Niederle, and Lise Vesterlund. 2020. "Knowing When to Ask: The Cost of Leaning In." *Journal of Political Economy* 128, no. 3 (March): 816–854.

Gagnon, Nickolas, Kristof Bosmans, and Arno Riedl. 2024. *The Effect of Gender Discrimination on Labor Supply.* SSRN Scholarly Paper. Rochester, NY, February.

Grossman, Philip J., Catherine Eckel, Mana Komai, and Wei Zhan. 2019. "It pays to be a man: Rewards for leaders in a coordination game." *Journal of Economic Behavior & Organization* 161 (May): 197–215.

Iriberri, Nagore, and Pedro Rey-Biel. 2021. "*Brave* boys and *play-it-safe* girls: Gender differences in willingness to guess in a large scale natural field experiment." *European Economic Review* 131 (January): 103603.

Koutout, Kristine. 2022. *Gendered Beliefs and the Job Application Decision: Evidence from a Large-Scale Field and Lab Experiment.* SSRN Scholarly Paper. Rochester, NY, February.

Leibbrandt, Andreas, and John A. List. 2015. "Do Women Avoid Salary Negotiations? Evidence from a Large-Scale Natural Field Experiment." *Management Science* 61 (9): 2016–2024.

Lepage, Louis-Pierre, Xiaomeng Li, and Basit Zafar. 2025. *Anticipated Discrimination and Major Choice.* Working Paper, April.

Li, Cher Hsuehhsiang, and Basit Zafar. 2023. "Ask and You Shall Receive? Gender Differences in Regrades in College." *American Economic Journal: Economic Policy* 15, no. 2 (May): 359–394.

Niederle, Muriel, and Lise Vesterlund. 2007. "Do Women Shy Away From Competition? Do Men Compete Too Much?*." *The Quarterly Journal of Economics* 122, no. 3 (August): 1067–1101.

Osun, Elif B. 2024. "Gender differences in advice giving." *Experimental Economics* (November).

Recalde, Maria, and Lise Vesterlund. 2020. *Gender Differences in Negotiation and Policy for Improvement.* Working Paper, December.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales. 2014. "How stereotypes impair women's careers in science." *Proceedings of the National Academy of Sciences* 111, no. 12 (March): 4403–4408.

Rheault, Ludovic, Erica Rayment, and Andreea Musulan. 2019. "Politicians in the line of fire: Incivility and the treatment of women on social media." *Research & Politics* 6, no. 1 (January): 2053168018816228.

Roberts, Tomi-Ann, and Susan Nolen-Hoeksema. 1989. "Sex differences in reactions to evaluative feedback." *Sex Roles* 21, no. 11 (December): 725–747.

Roussille, Nina. 2024. "The Role of the Ask Gap in Gender Pay Inequality." *The Quarterly Journal of Economics* 139, no. 3 (August): 1557–1610.

Rudman, Laurie A., Corinne A. Moss-Racusin, Julie E. Phelan, and Sanne Nauts. 2012. "Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders." *Journal of Experimental Social Psychology* 48, no. 1 (January): 165–179.

Ruebeck, Hannah. 2025. *Causes and Consequences of Perceived Workplace Discrimination.* SSRN Scholarly Paper. Rochester, NY, July.

Sarsons, Heather. 2022. *Interpreting Signals in the Labor Market: Evidence from Medical Referrals.* Working Paper, September.

Sarsons, Heather, Klarita Gërxhani, Ernesto Reuben, and Arthur Schram. 2021. "Gender Differences in Recognition for Group Work." *Journal of Political Economy* 129, no. 1 (January): 101–147.

Saygin, Perihan, and Thomas Knight. 2023. *Gender Bias in Peer Performance Evaluations: Evidence from a Field Experiment.* SSRN Scholarly Paper. Rochester, NY, January.

Saygin, Perihan O., and Ann Atwater. 2021. "Gender differences in leaving questions blank on high-stakes standardized tests." *Economics of Education Review* 84 (October): 102162.

Shastry, Gauri Kartini, and Olga Shurchkov. 2024. "Reject or revise: Gender differences in persistence and publishing in economics." *Economic Inquiry* 62 (3): 933–956.

Shastry, Gauri Kartini, Olga Shurchkov, and Lingjun Lotus Xia. 2020. "Luck or skill: How women and men react to noisy feedback." *Journal of Behavioral and Experimental Economics* 88 (October): 101592.

Small, Deborah A., Michele Gelfand, Linda Babcock, and Hilary Gettman. 2007. "Who goes to the bargaining table? The influence of gender and framing on the initiation of negotiation." *Journal of Personality and Social Psychology* 93 (4): 600–613.

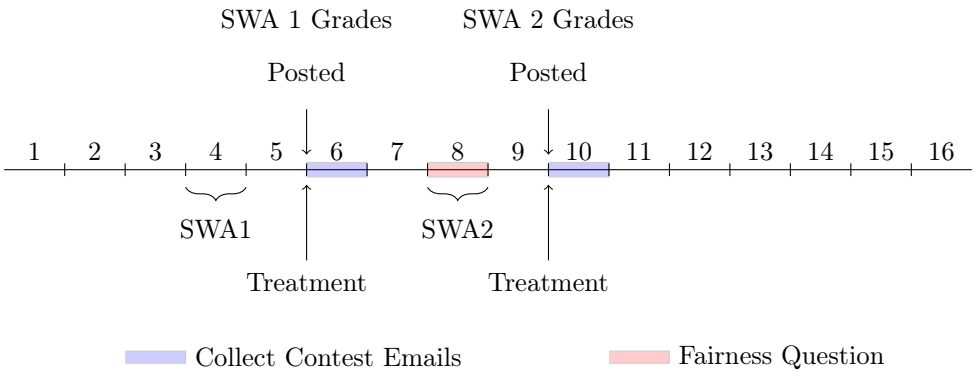
Ugalde Araya, Maria Paola. 2024. *Gender, Grade Sensitivity, and Major Choice.* Working Paper, November.

Wu, Alice H. 2018. "Gendered Language on the Economics Job Market Rumors Forum." *AEA Papers and Proceedings* 108 (May): 175–179.

Appendix

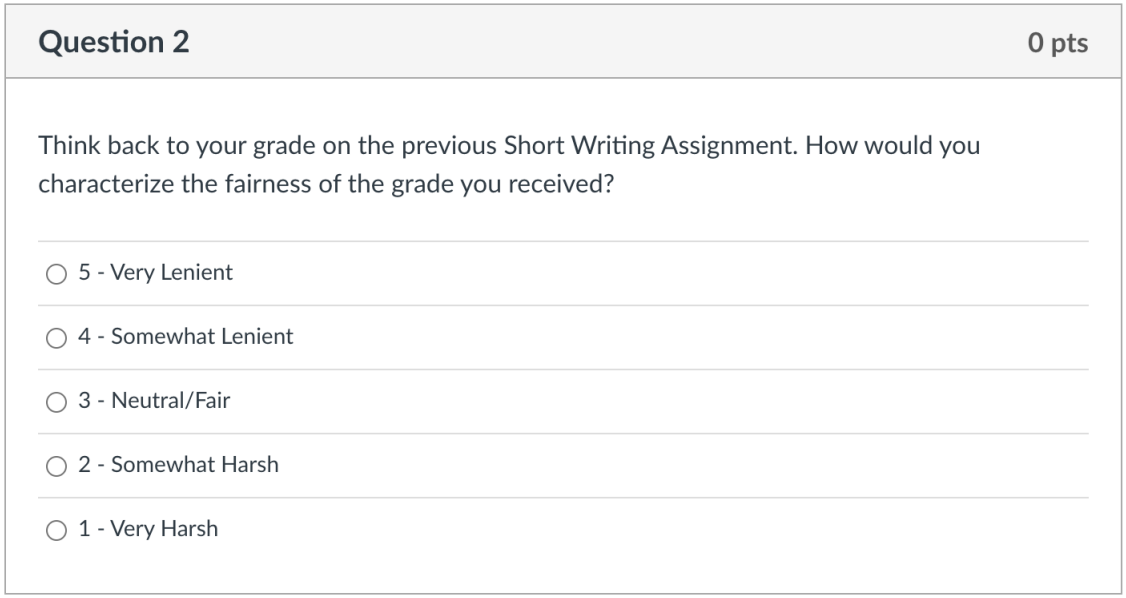
A. Experimental Design

Figure A1: Experiment Timeline



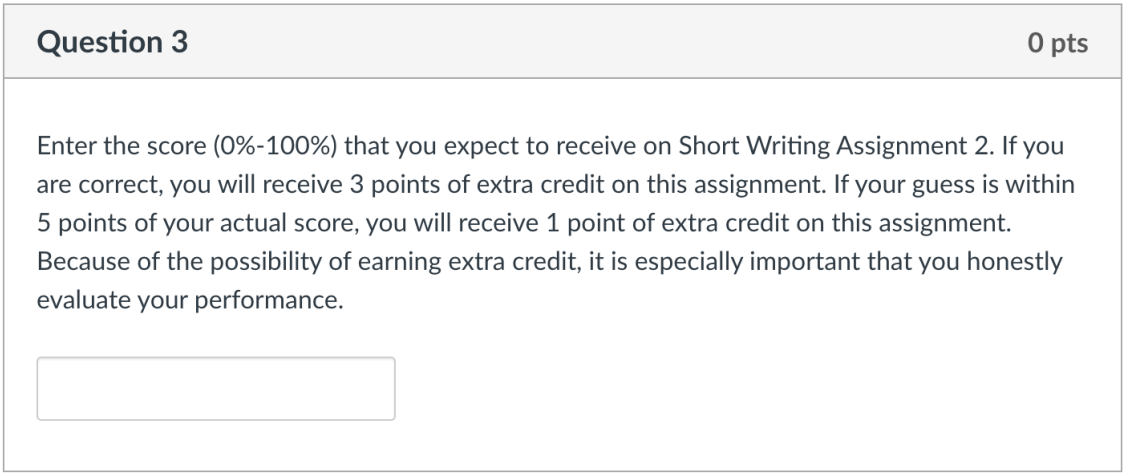
Notes. The figure presents a timeline of the typical 16-week semester and is not exact. The courses each consisted of weekly quizzes and well as three exams spaced every 5-6 weeks.

Figure A2: Fairness Question



Notes. The screenshot above is of the question asked before students submitted the second SWA in both courses across both semesters.

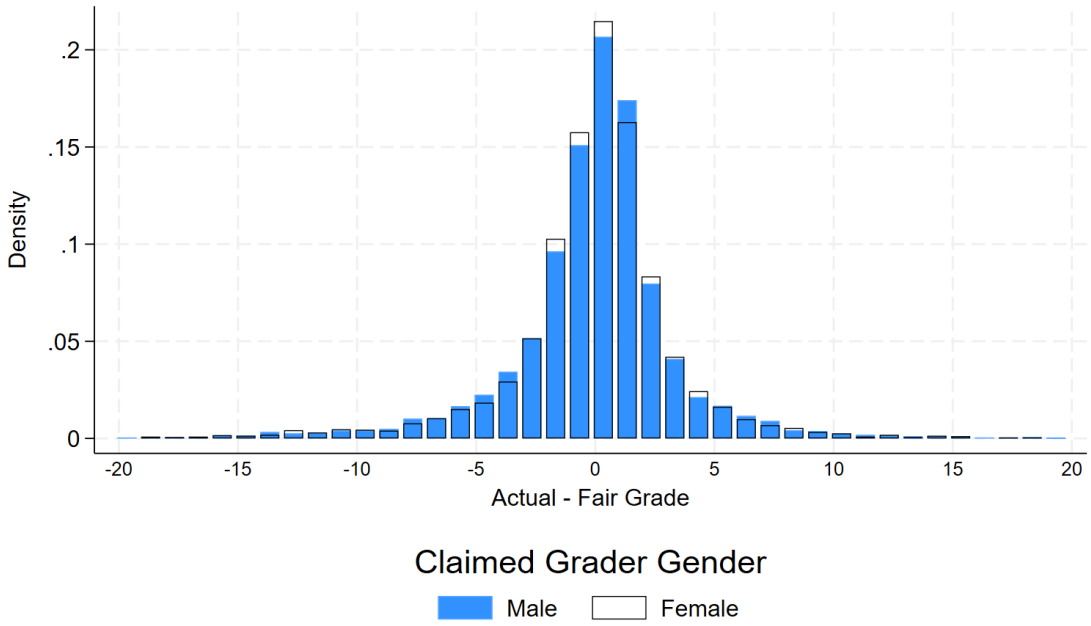
Figure A3: Score Prediction Question



Notes. The screenshot above is the language used in Spring 2024. In Fall 2023 we provided 5 points for an exact score prediction and 3 points for being within 5 points. The respective language was asked before students submitted either SWA in both courses.

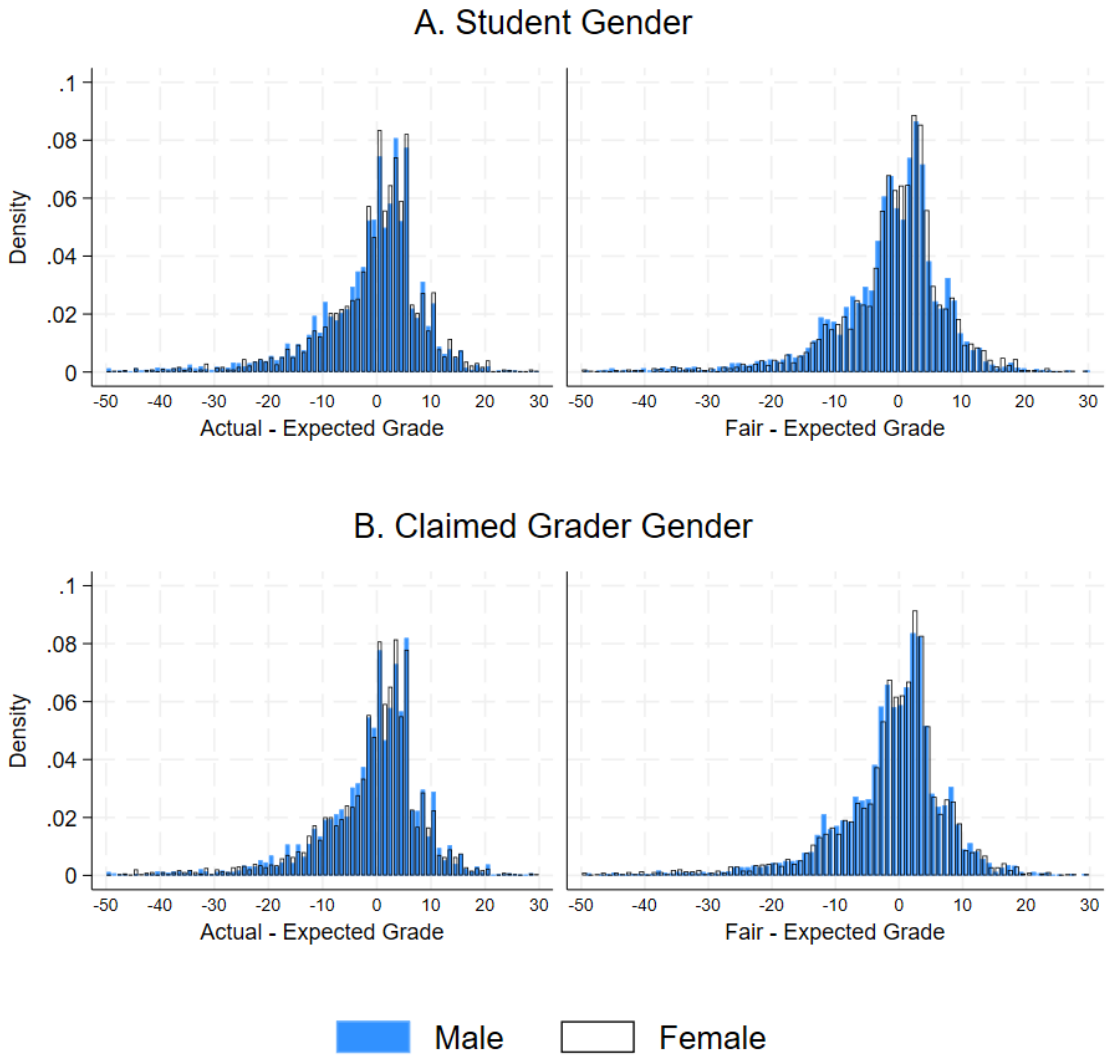
B. Additional Summary Statistics

Figure B1: Distribution of Lenience Measure



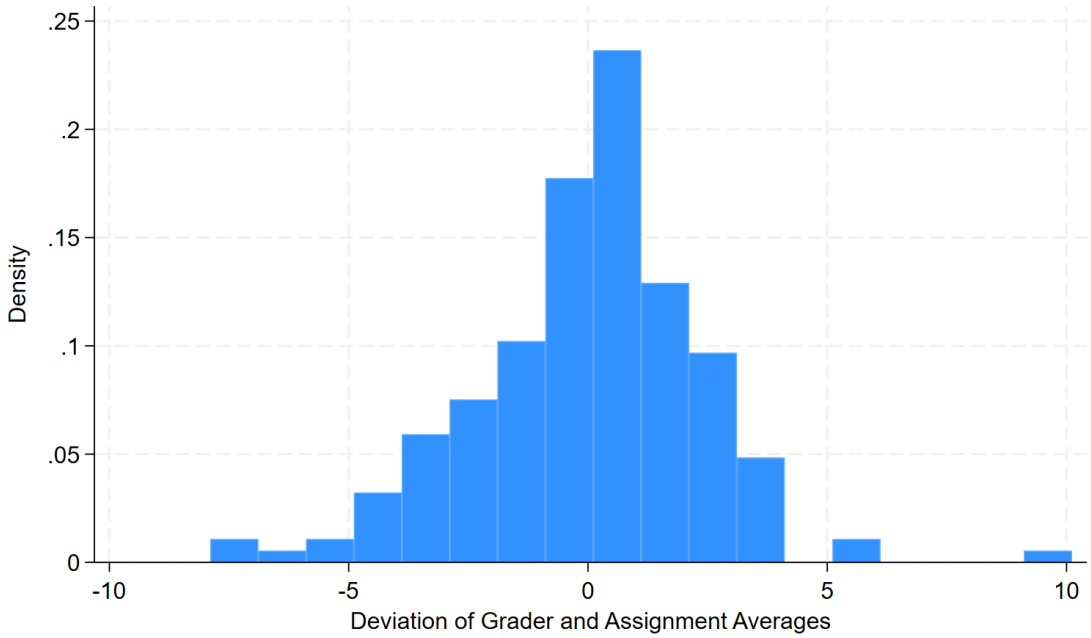
Notes. The sample excludes missing or improperly-submitted assignments. The figure is restricted to $|\text{Actual - Fair}| \leq 20$.

Figure B2: Distribution of Deviation of Scores from Students' Expected Scores



Notes. The sample excludes missing or improperly-submitted assignments. The figure is restricted to domain -50 to 30.

Figure B3: Distribution of Deviation of Grader and Assignment Average Scores



Notes. The histogram plots the distribution of the deviation between grader-specific and assignment average scores for the 186 grader-semester-course-assignment combinations. The sample excludes missing or improperly-submitted assignments.

Table B1: Summary Statistics with Sample Restrictions by Student Gender

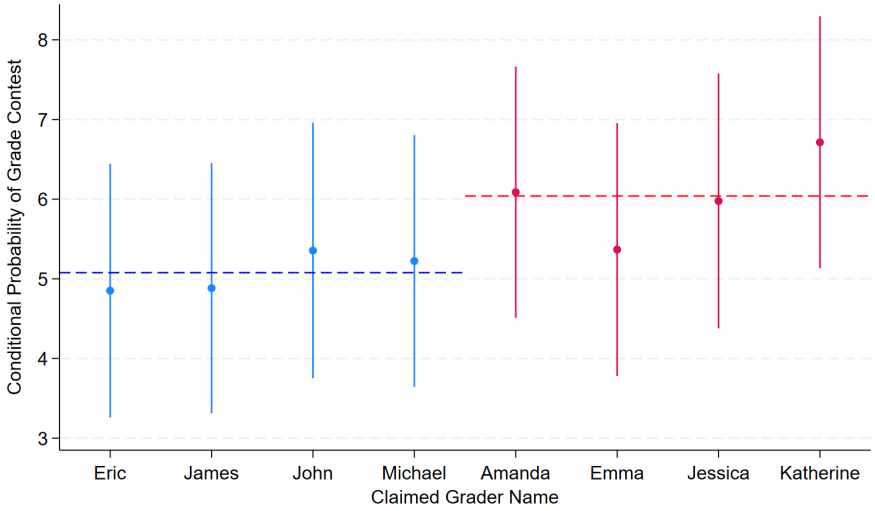
	(1)	(2)	(3)
Variable	Male Student	Female Student	Difference (se)
Female Claimed Grader	0.514	0.488	-0.026**
	(0.500)	(0.500)	(0.011)
Female Actual Grader	0.506	0.512	0.006
	(0.500)	(0.500)	(0.011)
Actual SWA Score	91.883	92.639	0.756***
	(10.960)	(10.236)	(0.230)
Fair Grade	91.911	92.596	0.685***
	(10.219)	(9.587)	(0.215)
Expected SWA Score	93.763	93.743	-0.020
	(6.005)	(6.066)	(0.131)
Actual - Fair Grade	-0.029	0.043	0.071
	(4.083)	(3.959)	(0.087)
Actual - Expected Grade	-1.881	-1.104	0.776***
	(10.847)	(10.333)	(0.230)
Fair - Expected Grade	-1.852	-1.147	0.705***
	(10.112)	(9.605)	(0.214)
SD of Regrades (Ambiguity)	2.010	2.044	0.034
	(2.957)	(3.307)	(0.068)
Contested Prior Grade	0.034	0.034	-0.001
	(0.182)	(0.180)	(0.004)
Course Grade after SWA2	80.500	81.324	0.824***
	(13.283)	(13.329)	(0.290)
Contested Grade to Instructor	5.119	5.208	0.089
	(22.040)	(22.222)	(0.481)
Contested Content	4.469	4.629	0.160
	(20.666)	(21.014)	(0.453)
Contested Writing	2.322	2.515	0.193
	(15.062)	(15.660)	(0.334)
Harsh	23.014	22.227	-0.787
	(42.103)	(41.587)	(1.297)
Observations	4,005	4,493	8,498

Notes: Unsubmitted and improperly submitted assignments (those receiving a 0 on writing or those regraded due to name appearing) and clear grader errors are omitted.

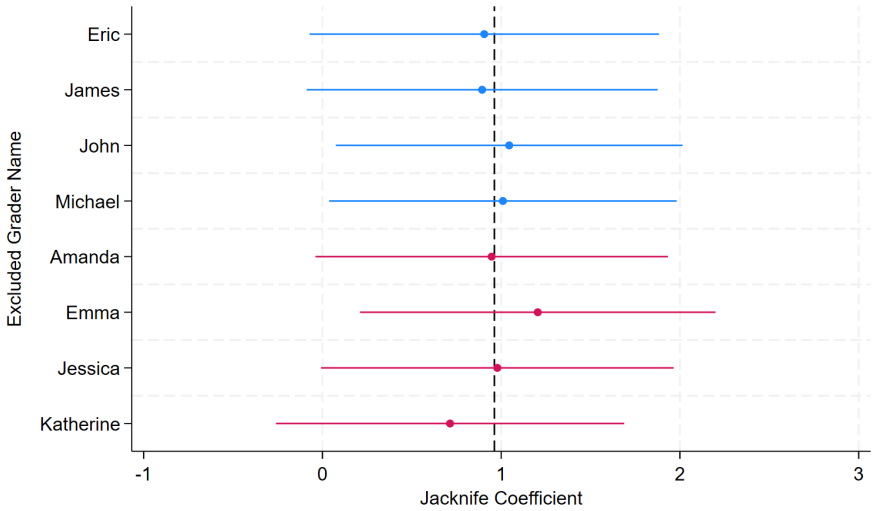
C. Robustness of Main Result

Figure C1: Robustness by Grader Name Inclusion

(a) Conditional Probabilities



(b) Jackknife Coefficients



Notes. Figure (a) plots the name-specific coefficients from the baseline regression specification where the female grader dummy is replaced with a vector of name-specific dummies. The mean of the male-specific and female-specific coefficients (5.078 and 6.040, respectively) are also plotted. Figure (b) plots the main coefficient of interest from a series of 8 regressions where in each a separate grader name is excluded. In all regression estimations the sample excludes missing or improperly-submitted assignments. 95% confidence intervals using robust standard errors are displayed as bands.

D. Fair Grade Calculation

We prefer to use a measure of the fair grade which includes the actual assigned score as it produces an estimator with a lower variance while remaining unbiased. To see this, let $X_i \overset{\text{iid}}{\sim} G(\mu)$ denote a random draw from a distribution of possible grades G, which has mean (fair-grade) μ. Let $R_1, R_2, ..., R_{n_r} \overset{\text{iid}}{\sim} G(\mu)$ denote n_r random draws from the same distribution. Define $L_1 = X_i - \frac{1}{n_r}\left(\sum_{i=1}^{n_r} R_i\right)$ and $L_2 = X_i - \frac{1}{n_r+1}\left(X_i + \sum_{i=1}^{n_r} R_i\right)$. Then,

$$
\begin{aligned}
Var(L_1) &> Var(L_2) \\
Var\left(X_i - \frac{1}{n_r}\sum_{i=1}^{n_r} R_i\right) &> Var\left(\frac{n_r X_i}{n_r+1} - \frac{1}{n_r+1}\sum_{i=1}^{n_r} R_i\right) \\
Var\left(X_i\right) + \frac{1}{n_r^2} Var\left(\sum_{i=1}^{n_r} R_i\right) &> \frac{n_r^2}{(n_r+1)^2} Var\left(X_i\right) + \frac{1}{(n_r+1)^2} Var\left(\sum_{i=1}^{n_r} R_i\right) \\
\frac{(n_r+1)^2 - n_r^2}{(n_r+1)^2}\left[Var\left(X_i\right) + Var\left(\sum_{i=1}^{n_r} R_i\right)\right] &> 0
\end{aligned}
$$

is true for any $n_r > 0$. Clearly both L_1 and L_2 are unbiased. Thus, L_2 is the preferred estimator.