Dondena Working Papers

Carlo F. Dondena Centre for Research
on Social Dynamics and Public Policy

Does Scarcity of Female Instructors Create Demand for Diversity among Students? Evidence from an M-Turk Experiment

Patricia Funk
Nagore Iriberri
Giulia Savio

Working Paper No. 148
September 2021

Università Bocconi • The Dondena Centre

Via Guglielmo Röntgen 1, 20136 Milan, Italy
http://www.dondena.unibocconi.it

ISSN-2035-2034

Does Scarcity of Female Instructors Create Demand for Diversity among Students? Evidence from an M-Turk Experiment*

Patricia Funk† Nagore Iriberri‡ Giulia Savio§

September 6, 2021

Abstract

Scarcity of female academics has been well documented for math-intensive or STEM fields. We investigate whether a lack of female instructors creates a demand for diversity on the student side. In an incentivized instructor-choice experiment on MTurk, we experimentally vary the gender balancedness of the instructor pool and let participants choose one additional instructor among one male and one female. We find that only women are more likely to choose the female instructor when the pool of instructors is male-dominated, suggesting that female students appreciate a more balanced instructor pool if female professors are scarce. We further document that women also appreciate diversity (though to a lesser extent) if the scarce gender is of the opposite sex. In contrast, men only appreciate diversity if the scarce gender is their own.

Keywords: instructor-choice experiment, gender scarcity

*We would like to thank Università della Svizzera italiana for providing anonymized data on students' teaching evaluations and choices on elective courses. We are indebted to Alejandro Martínez Marquina for his invaluable help with the implementation of the MTurk experiment. We also thank Muriel Niederle for her very helpful advice on the experimental design. Further thanks for helpful comments goes to Stefano DellaVigna, Anne Boring, Pedro Rey-Biel and seminar participants at Università della Svizzera italiana, Universitat Pompeu Fabra, University of Zürich, University of Milan and Erasmus University Rotterdam. Finally, we thank Silvia Martinez and Ludovica Marcotti for excellent research assistance. Patricia Funk acknowledges financial support from the Swiss National Science Foundation (grant no. 178887). Nagore Iriberri acknowledges financial support from Ministerio de Economía y Competición (PID2019-106146GB-I00), Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco (IT1367-19). All remaining errors are our own.

†Università della Svizzera italiana. E-mail: patricia.funk@usi.ch.

‡University of the Basque Country UPV/EHU, IKERBASQUE, Basque Foundation for Science. E-mail: nagore.iriberri@gmail.com.

§Università Bocconi. E-mail: giulia.savio@unibocconi.it.

1 Introduction

In the last 50 years women have made dramatic gains in science. Still, female representation is very unequal across different fields (Ceci and Williams, 2011, and Ceci et al., 2014). While in life sciences and some social sciences female representation has increased considerably, reaching or even surpassing parity, in the most math-intensive fields women's representation is still low. Economics is among the latter. According to the most recent survey by the American Economic Association, 23.5 percent of tenured and tenure-track faculty in economics are women. As such, gender diversity among economics academics is as poor as in the male-dominated tech industry, where 30 percent of the Silicon Valley workforce is female. Even worse, among full professors in economics, the share of women is often less than 15 percent (Lundberg and Stearns, 2019).

This scarcity of female economists has recently attracted considerable attention (Chari and Goldsmith-Pinkham, 2017; Lundberg and Stearns, 2019). While the point has been made that lack of women may have negative consequences for research (Bayer and Rouse, 2016), lesser thought has been given to potential negative effects on students. However, as a lack of diversity affects the type of research topics studied and taught to the students, this factor may directly channel into female and male students' interest in economics and other math-intensive fields. Moreover, teaching styles may vary with instructor gender and affect student satisfaction of either, or both sexes. In sum, if students value diversity in the instructor pool, a low share of female instructors may make them more valuable in the students' eyes. This taste for female instructors could be driven by all students (general taste for diversity), or among certain subgroups of students in particular. Concerning the latter channel, research in social psychology suggests that an individual's distinctive trait in relation to other people in the environment is more salient if this trait is a numerical minority ("numerical distinctiveness theory", see McGuire and Padawer-Singer, 1976; McGuire and McGuire, 1981). As such, when female professors are scarce, gender may become particularly salient to female students, which may affect their preferences for female (as opposed to male) instructors.

We directly test for the presence of a (potentially gender-specific) taste for instructor diversity in an experimental setting. With this aim, we design a deception-free, incentivized instructor-choice experiment on MTurk, where participants have to select an instructor whose advice they can read on how to solve a given task. The choice set consists of two instructors with comparable qualifications and experience but different gender. Before getting to this choice (which is our key outcome of interest), participants are told that there is a pool of six instructors, all of whose advice they can select to read. To test whether scarcity of women affects the choice of the additional instructor (male or female), we experimentally vary the "stock" of six instructors. In the balanced treatment, a participant is presented a stock of three female and three male instructors, whereas in the unbalanced treatment, the participant has a stock of six male instructors.

The main interest of the experiment is to analyze whether the choice of the additional instructor (male or female) depends on the gender balancedness of the existing instructor pool. To rule out that the order of presenting the two instructors (or details of their profile) affects the participants' choice, we randomize participants into permutations that vary according to task type, the order of presenting the two candidates, and the values of the two characteristics attached to the candidates (all details are provided in Section 2). To ensure that participants take the experiment seriously, we use a variable remuneration that increases with the correct answers in the tasks (next to a fixed show-up fee).

Our main findings are the following. First, scarcity in the stock of female instructors positively affects the probability of having a female chosen as the additional instructor. On average, the female instructor is 11 percentage points more likely to be selected if the stock of instructors is gender-unbalanced. Second, female and male participants react differently to scarcity in the instructor pool. If female instructors are scarce, female participants are 12.3 percentage points more likely than male participants to select the female instructor (this difference is highly statistically significant). Moreover, in contrast to female participants, male participants do not react to scarcity of female instructors in a statistically significant way.

While the experimental setting mimics the case of underrepresentation of women (as present in STEM and economics), two plausible mechanisms can explain the results. First, female participants prefer female instructors when female instructors are scarce. Second, female participants have a general preference for diversity (independent of whether scarcity refers to their own gender). To investigate the plausibility of the second channel, we compare instructor-choices of the gender-balanced treatment with a new unbalanced treatment, where all instructors are female. We find that women also value diversity when male instructors are scarce, but to a lesser extent than when female instructors are scarce. By contrast, men value diversity *only* if the scarcity is related to their own gender.

In summary, our experiment indicates that gender-related preferences emerge differently in different contexts. When women are scarce, they become more valuable, particularly among the subgroup of female decision makers. Taken at face value, our experiment implies that female students would ceteris paribus have a preference for female lecturers and professors when they are scarce. While students typically can not choose (let alone hire) lecturers and professors, we expect that a demand for diversity may show up in two areas: the choice of elective courses and student evaluations. In male-dominated faculties, female students may be more inclined to choose elective courses taught by female professors, compared to more gender-balanced faculties. Second, when courses cannot be chosen (compulsory courses), female students may appreciate being taught by a female professor, especially in fields where women are scarce - and this may show up positively in the teaching evaluations. Using data from a Swiss university, we find indeed evidence supporting these diversity claims. As such, we believe that increasing the share of women in male-dominated faculties (e.g., STEM disciplines) may increase student satisfaction and act as a pull-factor for future female students.

Our study contributes to four main strands of the literature. The first is on hiring women in academia. As studies relying on observational data are problematic because of unobserved quality differences between candidates, the most convincing studies exploit experimental variation. An early study by Steinpreis et al. (1999) studied a hypothetical hiring decision among psychology faculty, where the gender of the candidate was experimentally varied. The main finding was that both male and female faculty were less favorable towards the female candidate. More recently, Williams and Ceci (2015a) conducted a similar hypothetical hiring experiment among faculty in biology, engineering, economics, and psychology. Surprisingly, the results show a consistent preference for women, with the exception of male economists, who were found to be gender-neutral. Our main contribution to these papers is to causally identify the effect of scarcity in a setting where the hiring decision is incentivized. Our results support the view that, especially in settings where women are scarce, female candidates have an edge.

Second, we find that on average, female participants prefer female candidates, especially when female candidates are scarce. This result complements previous research on ingroup favoritism and outgroup bias (see Tajfel et al., 1971; Chen and Li, 2009; Chen and Chen, 2011; and Chen et al., 2014; and Coffman et al., 2018). We add to this strand of literature a connection between the strength of this ingroup preference and the scarcity of the ingroup. As becomes apparent from our study, ingroup preferences may become amplified when the ingroup gets relatively smaller.

Third, our paper relates to literature on diversity. While there is a literature exploring the consequences of diversity (Apesteguia et al., 2012; Hoogendoorn et al., 2013), our paper presents novel evidence on the demand for diversity.

Fourth, our paper relates to literature that documents gender differences in student evaluations. While female students in Economics appear to be more critical than males when evaluating male professors, the same does not hold when evaluating female professors (Boring, A., 2017; Mengel et al., 2019). We replicate a same-sex preference of female students in Economics, but also document that gender differences in instructor evaluations are completely absent in more gender-balanced faculties (Communication), and get even aggravated in more unbalanced faculties (Computer Science).

The remainder of this article is structured as follows. Section 2 describes the experimental design and shows the main results. Section 3 complements the results from the experiment with observations from university data, adding external validity to the experimental findings. Section 4 concludes.

2 Field Experiment on MTurk

2.1 Design and Data

We designed an incentivized and deception-free instructor-choice experiment on MTurk. Participants choose an instructor to receive advice on how to solve a given task under time pressure.

Participants are randomized into two types of tasks: mathematical multiplications ("math task") or spelling certain English words correctly ("English task"). At the beginning of the experiment, participants are given information on the payoff structure (payoff depends on task performance) and the type of task (math or English, without giving any further details). All participants are informed that they will receive 1 dollar for their participation plus 40 cents for each correct answer. Most importantly for this experiment, participants are told that they can read tips on how to solve the tasks (see Appendix for details) by six instructors (selected by us). In addition, they can choose *one additional instructor*, whose advice they can read. In the end, they will have to choose among one male and one female instructor with comparable qualifications and experience. The key feature of this experiment is that we experimentally vary the "stock" of six instructors. In the balanced treatment, the participant has a stock of three female and three male instructors, whereas in the unbalanced treatment, the participant has a stock of six male instructors. The information given to the participants is the instructor's name (Margaret or Richard), the fact that he/she is a graduate student, the GPA (3.5 or 3.6 out of 4), and the accumulated hours as a teaching assistant (29 or 31). See Figure 1 for a screenshot for the treatment (unbalanced) and control (balanced).[1] The main interest of the experiment is analyzing whether the choice of Margaret (as opposed to Richard) as an additional instructor depends on the treatment.

We design 16 permutations, 8 for the math task and 8 for the English task. For each task type, 4 permutations have a balanced instructor pool and 4 permutations have an unbalanced instructor pool. These 4 permutations differ in the order of instructor presentation (Margaret first or second) and characteristics (Margaret with a higher GPA but fewer accumulated hours as TA or Margaret with a lower GPA but more accumulated hours as TA). The goal was to obtain roughly 100 participants for each permutation, leading to a total of 1,600 participants. We managed to collect 1,955 observations. However, we removed all participants who tried to run the experiment twice and those who appeared to be doing the experiment together with a second person.[2] This left us with a participant pool of 1,478: summary statistics are reported in Table 1. As shown in Panel A, randomization of CV-characteristics (GPA and hours of experience as TA) across the two candidates' profiles worked well, as the likelihood that Margaret comes first or that Margaret has a higher GPA is always approximately 50%. As evident from Panel B, all demographic covariates are balanced across treatments. The typical participant is white, in possession of a college degree, and in his/her mid-thirties. The share of female participants is slightly below 50 percent and comparable across balanced and unbalanced instructor pools. In Panel C, we report participants' behavior during the experiment. As expected, the main endogenous variable of interest (instructor choice) differs across treatments. Margaret is chosen more frequently when the treatment is "Unbalanced" (when female instructors are scarce). Regarding the duration of the task, the number of times instructor advice was sought, or performance in terms of correct answers, we do not see any differences across treatments.

[1]In the balanced treatment, the participants are told that they have six instructors "Jim", "Mary", "John", "Patricia", "Robert" and "Linda", all graduate students. In the unbalanced treatment, the participants are told that they have six instructors "Jim", "Kevin", "John", "William", "Robert" and "David", all graduate students. The actual tips are obtained from real graduate students who were shown the task and were asked to describe the task in written form.

[2]That is, two participants running the experiment with the same IP address.

While the summary statistics indicate that the participants check for advice slightly more than 4 times on average, it is also interesting to look at *which* type of advice the participants seek. In Figure 2, we document the percentage of participants who click on a specific advice, starting with advice from the instructor to the farthest left of the instructor pool (Tip 1, referring to the advice from Jim), followed by advice from the second-leftmost instructor (Tip 2, referring to advice from Kevin in the unbalanced treatment and Mary in the balanced treatment), etc. The advice number 7 (Tip 7) is the advice from the instructor chosen by the participant (Margaret or Richard). As shown in Figure 2, a spike is observed for Tip 7 (for both male and female participants), meaning that advice is most frequently sought from the participant-selected instructor.

Of 1,478 participants, only 267 did not look at the advice of their chosen instructor. We present the main results for the 1,009 participants who actually looked at the advice of their selected instructor. Arguably, these participants took the instructor-choice decision most seriously, as they did (and likely planned to) look at the instructor's advice. Moreover, we will also document the robustness to alternative data samples.

2.2 Main Results: Instructor Choice in the Presence of Female Scarcity

To identify the causal effect of scarcity of women on demand for diversity, we run two regression equations:

$$Margaret_i = \alpha + \beta Unbalanced_i + \eta Math_i + \theta MargFirst_i + \psi MargTA_i + \iota' X_i + \epsilon_i \quad (1)$$

$$Margaret_i = \alpha + \beta Unbalanced_i + \gamma Female_i \times Unbalanced_i + \eta Math_i + \theta MargFirst_i + \psi MargTA_i + \iota' X_i + \epsilon_i \quad (2)$$

The dependent variable $Margaret_i$ is a dummy equal to one if participant i chooses the female candidate (Margaret) over the male candidate (Richard). $Female_i$ is a dummy equal to one if the participant is female. $Unbalanced_i$ is a dummy equal to one if participant i is exposed to a pool of six male instructors. The variables $Math_i$, $MargFirst_i$ and $MargTA_i$ control for the experimental permutation: $Math_i$ is a dummy equal to one for the math task; $MargFirst_i$ is a dummy equal to one if - in the instructor choice step - the name of the female candidate (Margaret) comes before the name of the male candidate (Richard); and $MargTA_i$ is a dummy variable taking a value of one if the female candidate (Margaret) is more experienced as a teaching assistant than is the male candidate (Richard). Finally, X_i is a vector of individual covariates listed in Table 1.

The main coefficient of interest in equation 1 is β. A positive β suggests that female instructors are more frequently selected when scarce. Equation 2 adds an interaction term $Female \times Unbalanced$, γ, which enables testing for whether the potential effect of (female) scarcity on instructor choice is gender-specific.

Our main experimental results are shown in Table 2. Columns 1-2 show the results of regression equation 1. Clearly, being exposed to a pool of male instructors increases preferences for the female instructor. The probability of choosing Margaret increases by 11 percentage points if a participant is exposed to the gender-unbalanced instructor pool. Adding controls (column 2) hardly affects the estimated coefficient of the treatment "Unbalanced", as should be the case in successful randomization. Column 3 displays the results for regression equation 2, indicating that the stronger preference for female instructors in the treatment "Unbalanced" is *entirely* driven by female participants. Men are *not* more likely to choose Margaret if the teacher pool is unbalanced (see the estimated β).[3] These results suggests that women (but not men) value the diversity brought in by a female instructor if the pool of instructors is all male.

We perform a series of robustness tests, shown in Table 3. First, female participants may expect a different type of advice from female instructors that would help them to answer the questions correctly and earn more money. In this instrumental view, women select the female instructor because they would like to receive the advice. Alternatively,

[3] Note that we also estimated models with triple interaction terms to see whether effects differ between task type (English or math). Since the estimated coefficient before the triple interaction $Unbalanced \times Female \times Math$ is statistically insignificant, we report results for the two tasks combined.

the decision could be entirely unconscious, where women select the female instructor (when the instructor pool is all male), even though they have no expectations in terms of the advice they would receive from the female instructor. To test for this possibility, we conduct a placebo analysis. We restrict the sample to those participants who did not check any single advice (remember - before seeing the task!). These are likely participants whose strategy is to earn the participation fee but have no intention of exerting any additional effort to answer the questions correctly. As shown in Table 3 columns 1 and 2, there is no effect of the treatment "Unbalanced" (nor the interaction of "Unbalanced" with participant gender) on the probability that Margaret is chosen. As such, in the sample of participants who exert very little effort in the experiment, scarcity of women in the instructor pool does not affect the probability that the female instructor is chosen.

Second, most interesting to us are the participants who exert at least some minimal effort to correctly answer the questions. We presented the results for participants who looked at the advice of the chosen instructor in Table 2. As additional evidence, we now report the results for different participant samples, varying in the number of advice seen. As shown in Table 3, columns 3-8, gender-specific preferences for diversity get larger in samples where participants ask for more advice and, arguably, take the task more seriously. Last, we focus on those participants who could answer the last survey question: "How many women were in the initial instructor pool of six instructors?" Again, the effect is strong (in fact the strongest) among those participants who appeared to pay close attention to the experiment.

2.3 Additional Results: Instructor Choice in the Presence of Male Scarcity

We found that female participants value female instructors more when female instructors are scarce. While we were most interested in the setting lacking female diversity (which is the case in Economics and STEM fields), the question remains whether women value diversity *per se*, or only when scarcity refers to their own gender.

To get at the mechanism behind the previous results, we ran an additional treatment, where the scarce group is now the male one (i.e., the unbalanced treatment is all female). The rest of the design features (task type, the order of presenting the two candidates, and the values of the two characteristics attached to the candidates) are kept exactly in the same way as in the main experiment. Summary statistics for the scarcity of male instructors are presented in Table 4. Randomization worked well with a few exceptions regarding the socio-demographic variables (which we will control for in the regressions).

Table 5 shows the results. The dependent variable $Richard_i$ is a dummy equal to one if a participant i chooses the male candidate (Richard) over the female candidate (Margaret). As can be seen from columns 1 and 2, being exposed to a pool of female instructors increases preferences for the male instructor by about 13 percentage points. The magnitudes are therefore similar in both the female and male scarcity treatments (see Table 2, columns 1 and 2). In addition, in column 3, we show the potentially differential preference of female participants for the male instructor. As can be seen therefrom, the interaction term is negative, but not statistically significant. Therefore, women also value diversity in the scenario of male scarcity, but to a smaller extent compared to scarcity of their own gender (compare the estimated $(\beta+\gamma)$ with those in Table 2). By contrast, men value diversity *only* when male instructors are underrepresented - The estimated β is large and highly significant in Table 5, but not in Table 2.

In sum, we find that women value gender diversity in both cases, when female as well as male instructors are scarce. Men, in contrast, value diversity only when scarcity refers to their own gender.

3 External Validity: Observations from University Data

To what extent is our experiment informative about scarcity of women in academia? One direct implication of the experimental results is that in case of female professor scarcity, female students would opt to increase the female professor share. We cannot randomly vary the share of females in academia to test this prediction, and neither can we let students hire professors. However, using real-world university data, we expect the postulated taste for diversity to show up in two relevant student choices: the choice of elective courses, and the teaching evaluations of professors. We obtained data on elective choices and teaching evaluations for three faculties from the Università della Svizzera italiana (USI), which differ considerably in their scarcity of women (in increasing order): Communication, Economics, and Computer Science.

First, we hypothesize that the scarcer the female faculty, the more likely female students will choose elective courses taught by women. Do we see this hypothesized pattern reflected in students' choices of elective courses? We received information on all the exams students were taking in any elective course between 2015 and 2020. As some students may choose a course but not take the exam, this proxy for elective choice may contain a bit or measurement error. As the Bachelor in this university is quite structured and leaves little room for electives (88% of the courses are compulsory), we focus on exams taken at the Master level. Dropping duplicates of students who repeated an exam, we have 2,961 (student X exam) observations for Communication, 4,145 observations for Economics, and 861 observations for Computer Science. Of those, 1,905 are coming from female students in Communication, 2,107 in Economics, and 175 in Computer Science.

Figure 3 upper part shows the share of elective courses taught by female professors, for the three faculties.[4] The share of female professors in Communication is exactly the double compared to Computer Science, while Economics lies in between. Figure 3 lower

[4]The teaching staff includes professors and lecturers. To simplify language, I refer to both by professors.

part shows the surplus of female students in courses taught by female professors (in percent, relative to the share of female students in courses taught by male professors). As can be seen therefrom, the share of female students in Computer Science is about 20 percent higher when the course is taught by a female professor. Consistent with the finding that scarcity creates demand for diversity, this share is lower in Economics, and even more so in Communication. It is important to note that this is only suggestive evidence, as female professors may be teaching courses that are more appealing to female students.

Second, we hypothesize that in a situation where courses are exogenously given (as is the case in compulsory courses), female students appreciate female professors, and more so the more male-dominated the scientific field is. Do we see this reflected in the teaching evaluations? We collected teaching evaluations for all courses taught by the three faculties for the consecutive academic years of 2015-2016 and 2016-2017, and drop all elective courses.[5] In the academic year 2017-2018, a new evaluation system was introduced, so the newer data were no longer comparable. Before the academic year 2017-2018, teaching evaluations were done online after the students had taken the courses and completed the exams, but before they knew their actual grade. As filling out the teaching evaluations was necessary to access the grades, the response rate was close to 100%.[6] The teaching evaluation questionnaire consisted of 10 questions. We focus on the question that represents the summary evaluation of the course: "Please express your overall satisfaction with this course" (ranging from 1 (minimum) to 10 (maximum)).

The advantage of analyzing course evaluations is that - in contrast to the analysis on elective choice - we can partial out the course content, by running regressions with *course-fixed effects*. This means that we are effectively comparing evaluations for the *same course*. We run regressions, where the dependent variable is the teaching evalua-

[5] As can be seen from the summary statistics in Appendix Table A.1, more than one-half of the courses taught in 2015-2016 and 2016-2017 are compulsory.

[6] Under the new evaluation system, students no longer needed to fill out the course evaluations in order to see their grades, which led to a drop in the response rate. The high response rate under the old system led us to focus on the earlier data.

tion score given by student s to professor p teaching course c. Table 6 presents results where baseline estimates with course-year fixed effects are presented in columns 1, 4 and 7. In columns 2, 5, and 8, we add student fixed effects. Last, we add professor-course fixed effects that vary by year to account for the fact that some courses are co-taught (see columns 3, 6, and 9).

We focus on the interaction term ($Female_s X Female_p$), which tells us how the gender gap in the evaluations changes, when we move from the evaluation of male professors to the evaluation of female professors. While the differences-in-differences estimate is zero for Communication, it becomes positive for Economist, and even larger (albeit statistically insignifcant) for Computer Science. This is supportive evidence that female students in masculine faculties appreciate having a female professor.

4 Conclusions

Female underrepresentation in science (especially STEM faculties) is a topic of heated debate. While numerous articles explore potential causes (e.g., stereotypes (Reuben, 2014), family and career incompatibilities (Goldin, 2014), or publishing hurdles (Hengel, 2018; Card et al., 2020)), little is known about the consequences of a lack of academic diversity on students. In an incentivized and deception-free field experiment, we test how male and female participants value gender diversity in the instructor pool. When female instructors are scarce, we find that only female participants are more likely to choose an additional female instructor, when given the choice. As such, lack of female instructors generates a taste for diversity among females, but not males. On the other hand, when male instructors are scarce, both male and female participants value diversity, as shown by a higher likelihood of selecting another male instructor.

What are the implications for STEM faculties? Would female students be happier if more women were present? While we cannot directly test this with observational data, we provide two pieces of evidence that this may be the case: data from a Swiss university show that female students are more likely to select elective courses taught by female professors, if female professors are scarce. Second, in compulsory courses, female students show a more positive evaluation of female professors, the scarcer they are in the faculty. As such, in the most masculine faculties, female students seem indeed deprived of the diversity brought in by female instructors. Luckily for the few existing female students in STEM faculties, hiring preferences seem to become more female friendly as long as female candidates are equal to or better than male candidates (Williams and Ceci, 2015a, 2015b).

References

[1] Apesteguia, J., Azmat, G., and Iriberri, N. (2012). The impact of gender composition on team performance and decision making: Evidence from the field. *Management Science*, 58(1), 78-93.

[2] Bayer, A., and Rouse, C. E. (2016). Diversity in the economics profession: A new attack on an old problem. *Journal of Economic Perspectives*, 30(4), 221-42.

[3] Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, 145, 27-41.

[4] Card, D., DellaVigna, S., Funk, P., and Iriberri, N. (2020). Are referees and editors in economics gender neutral?. *The Quarterly Journal of Economics*, 135(1), 269-327.

[5] Ceci, S. J., and Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences*, 108(8), 3157-3162.

[6] Ceci, S. J., Ginther, D. K., Kahn, S., and Williams, W. M. (2014). Women in academic science: A changing landscape. Psychological science in the public interest, 15(3), 75-141.

[7] Chari, A., and Goldsmith-Pinkham, P. (2017). Gender representation in economics across topics and time: Evidence from the NBER summer institute (No. w23953). *National Bureau of Economic Research*.

[8] Chen, R., and Chen, Y. (2011). The potential of social identity for equilibrium selection. *American Economic Review*, 101(6), 2562-89.

[9] Chen, Y., and Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, 99(1), 431-57.

[10] Chen, Y., Li, S. X., Liu, T. X., and Shih, M. (2014). Which hat to wear? Impact of natural identities on coordination and cooperation. *Games and Economic Behavior*, 84, 58-86.

[11] Coffman, K. B., Exley, C. L., and Niederle, M. (2017). When gender discrimination is not about gender (No. 18-054). *Harvard Business School Working Paper.*

[12] "Women in Economics" (2017). *The Economist.*

[13] Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review*, 104(4), 1091-1119.

[14] Hengel, E. (2018) Publishing While Female: Are Women Held to Higher Standards? Evidence from Peer Review, University of Liverpool Working Paper.

[15] Hoogendoorn, S., Oosterbeek, H., and Van Praag, M. (2013). The impact of gender diversity on the performance of business teams: Evidence from a field experiment. *Management Science*, 59(7), 1514-1528.

[16] Lundberg, S. and Stearns, J. (2019). Women in Economics: Stalled Progress. *Journal of Economic Perspectives* 33(1), 3-22.

[17] McGuire, W. J., and Padawer-Singer, A. (1976). Trait salience in the spontaneous self-concept. *Journal of personality and social psychology*, 33(6), 743.

[18] McGuire, W. J., and McGuire, C. V. (1981). The spontaneous self-concept as affected by personal distinctiveness. *Self-concept: Advances in theory and research*, 147-171.

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

[20] Reuben, E., Sapienza P., and Zingales L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences* 111: 44034408.

[21] Steinpreis, R. E., Anders, K. A., and Ritzke, D. (1999). The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex roles*, 41(7-8), 509-528.

[22] Tajfel, H., Billig, M. G., Bundy, R. P., and Flament, C. (1971). Social categorization and intergroup behaviour. *European journal of social psychology*, 1(2), 149-178.

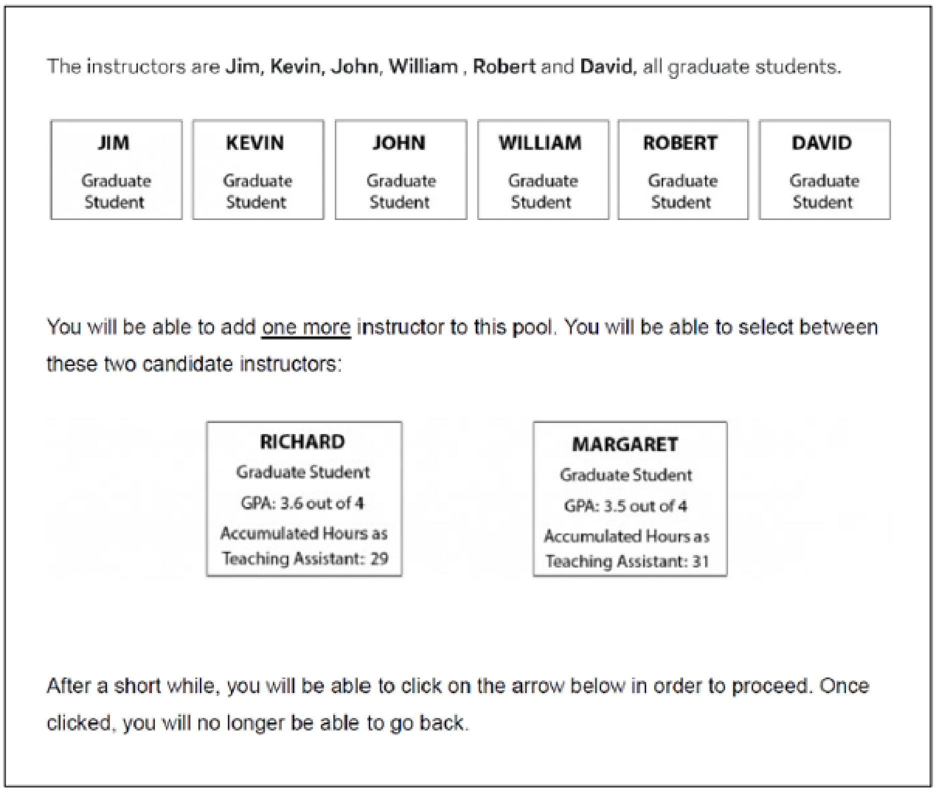
[23] Williams, W. M., and Ceci, S. J. (2015a). National hiring experiments reveal 2: 1 faculty preference for women on STEM tenure track. *Proceedings of the National Academy of Sciences*, 112 (17) 5360-5365.

[24] Williams, W. M., and Ceci, S. J. (2015b). Women have substantial advantage in STEM faculty hiring, except when competing against more-accomplished men *Frontiers in Psychology*, 6: 1532.

Figures and Tables

Figure 1: Experimental Treatments

Panel A: Treatment Unbalanced



Panel B: Treatment Balanced

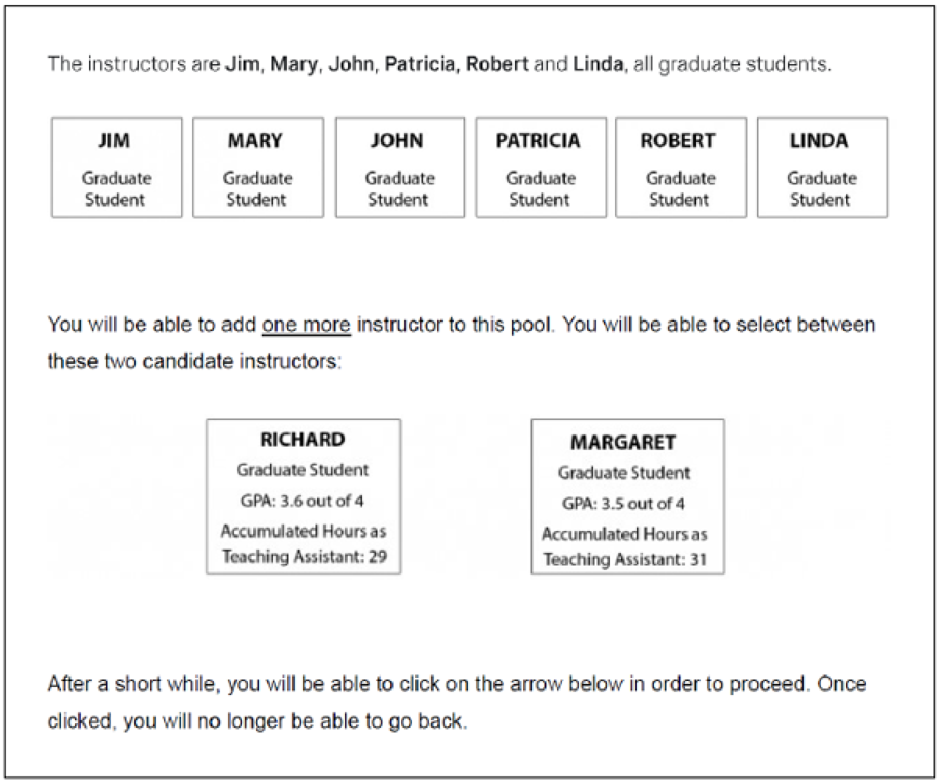


Figure 2: Percentage of Participants Checking Each Advice

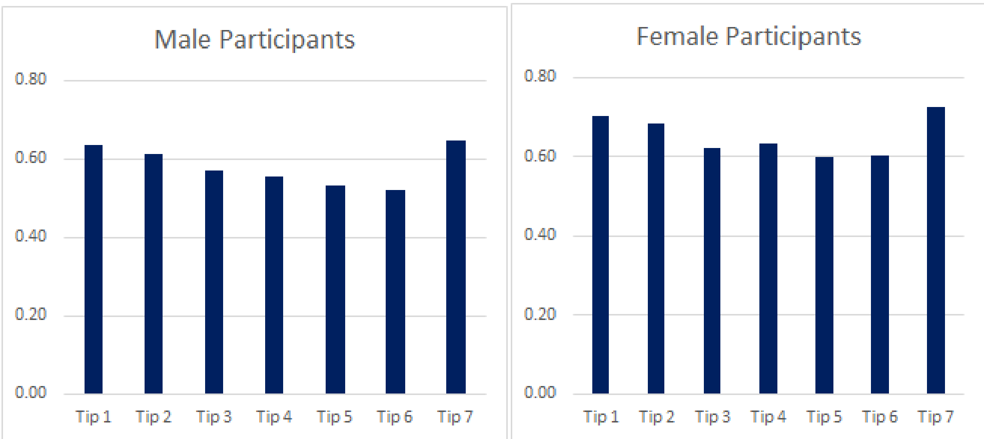


Figure 3: Elective Choices

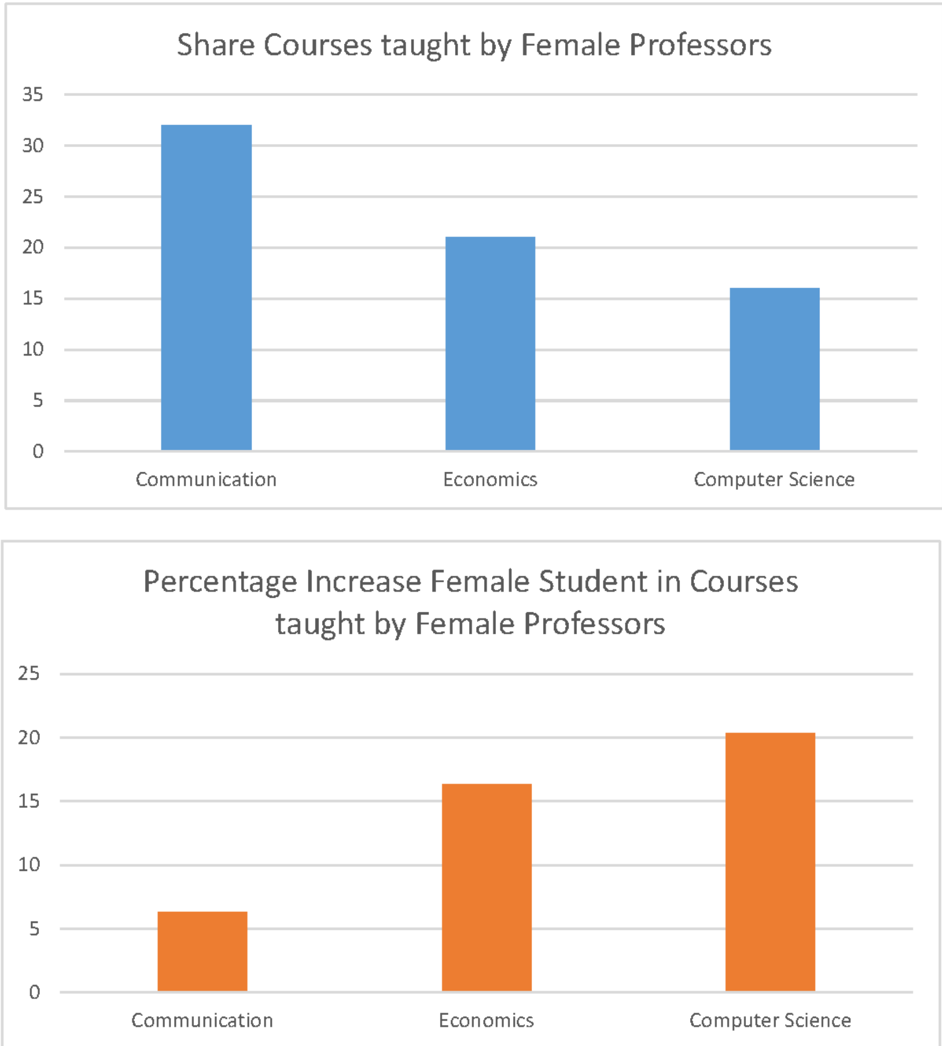


Table 1: Summary Statistics of MTurk Experiment: Balanced versus Unbalanced (Female Scarce)

Group	Balanced			Unbalanced			(B-U)
	No.Obs	Mean	Std.Dev	No. Obs	Mean	Std.Dev	P-value
Panel A: Permutation variables							
Math Task	743	0.47	0.50	735	0.47	0.50	0.886
Margaret First	743	0.49	0.50	735	0.50	0.50	0.756
Margaret TA	743	0.49	0.50	735	0.52	0.50	0.404
Panel B: Sociodemographic variables							
Female	743	0.45	0.50	735	0.48	0.50	0.18
Age	743	35.78	11.32	735	36.36	11.41	0.33
White	743	0.77	0.42	735	0.76	0.43	0.76
College degree	743	0.60	0.49	735	0.61	0.49	0.59
Post-graduate degree	743	0.30	0.46	735	0.31	0.46	0.75
Panel C: Participants' performance							
Margaret chosen	743	0.63	0.48	735	0.69	0.46	0.013
Duration	743	819.90	352.43	735	840.87	502.71	0.353
No. of advices	743	4.35	2.70	735	4.25	2.77	0.472
No. of correct answers	743	7.03	3.48	735	6.97	3.36	0.732

Notes. The group "Balanced" includes all participants exposed to a gender balanced pool of instructors, while the group "Unbalanced" includes all participants exposed to a pool of six male instructors. For each variable of interest, we report the number of observations, mean and standard deviation. The last column reports P-values of a t-test between variables in control and treatment group.

Table 2: Choice of Female Instructor when Female Instructors are Scarce

	(1)	(2)	(3)
Unbalanced (β)	0.116***	0.111***	0.049
	(0.029)	(0.029)	(0.041)
Female×Unbalanced (γ)			0.125**
			(0.053)
Math Task	0.006	0.009	0.006
	(0.025)	(0.024)	(0.024)
Margaret First	-0.055**	-0.059**	-0.059**
	(0.026)	(0.026)	(0.025)
Margaret TA	-0.022	-0.023	-0.021
	(0.038)	(0.039)	(0.038)
Female		0.050**	-0.014
		(0.023)	(0.036)
Age		0.004***	0.004***
		(0.001)	(0.001)
White		0.009	0.010
		(0.042)	(0.042)
College Degree		0.021	0.016
		(0.047)	(0.046)
Post-graduate Degree		0.013	0.005
		(0.051)	(0.050)
Constant	0.670***	0.494***	0.531***
	(0.038)	(0.069)	(0.067)
$\beta + \gamma$			0.174***
			(0.035)
R-squared	0.019	0.032	0.036
N	1009	1009	1009

Notes. The dependent variable is a dummy equal to one if Margaret is chosen. Treatment "Unbalanced" is a dummy equal to one if the participant is exposed to a pool of six male instructors, and zero if he/she is exposed to a gender balanced pool of instructors. All included participants checked the advice by the chosen instructor. Robust standard errors are reported in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Choice of Female Instructor when Female Instructors are Scarce: Robustness Checks

	Zero advices		One advice		Two advices		Three advices		Guessed right	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Unbalanced	-0.039	0.040	0.081***	0.057	0.098***	0.058	0.090***	0.028	0.110***	0.039
	(0.082)	(0.096)	(0.023)	(0.039)	(0.024)	(0.041)	(0.024)	(0.041)	(0.038)	(0.055)
Female×Unbalanced		-0.174		0.050		0.085		0.127**		0.128*
		(0.146)		(0.059)		(0.060)		(0.060)		(0.075)
Math Treatment	-0.048	-0.049	0.016	0.015	0.010	0.007	0.005	0.002	0.010	0.007
	(0.083)	(0.083)	(0.022)	(0.022)	(0.025)	(0.025)	(0.024)	(0.024)	(0.032)	(0.032)
Margaret First	0.062	0.061	-0.050**	-0.050**	-0.049*	-0.050*	-0.058**	-0.058**	-0.023	-0.026
	(0.074)	(0.074)	(0.022)	(0.022)	(0.026)	(0.025)	(0.028)	(0.028)	(0.037)	(0.037)
Margaret TA	0.002	0.016	-0.056*	-0.056*	-0.063*	-0.061*	-0.057*	-0.055	-0.083**	-0.079*
	(0.076)	(0.076)	(0.033)	(0.032)	(0.032)	(0.032)	(0.033)	(0.033)	(0.041)	(0.041)
Female	0.065	0.160**	0.067***	0.042	0.057**	0.016	0.057**	-0.005	0.025	-0.029
	(0.079)	(0.071)	(0.023)	(0.034)	(0.023)	(0.037)	(0.024)	(0.041)	(0.038)	(0.039)
Age	0.003	0.002	0.003***	0.003***	0.002**	0.002**	0.002*	0.002*	0.004***	0.004***
	(0.003)	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
White	-0.089	-0.081	0.033	0.033	0.042	0.042	0.029	0.030	0.075*	0.073
	(0.100)	(0.099)	(0.042)	(0.042)	(0.048)	(0.048)	(0.043)	(0.043)	(0.044)	(0.045)
College Degree	0.113	0.096	0.051	0.049	0.053	0.050	0.034	0.028	-0.014	-0.022
	(0.140)	(0.136)	(0.050)	(0.050)	(0.050)	(0.049)	(0.050)	(0.049)	(0.066)	(0.064)
Post-graduate Degree	0.060	0.037	0.033	0.031	0.046	0.042	0.036	0.029	0.024	0.010
	(0.153)	(0.146)	(0.052)	(0.052)	(0.052)	(0.052)	(0.052)	(0.052)	(0.068)	(0.065)
Constant	0.474***	0.458***	0.459***	0.472***	0.480***	0.505***	0.523***	0.562***	0.482***	0.524***
	(0.148)	(0.151)	(0.071)	(0.071)	(0.071)	(0.067)	(0.073)	(0.071)	(0.092)	(0.085)
R-squared	0.030	0.037	0.029	0.030	0.028	0.030	0.025	0.030	0.042	0.047
N	202	202	1276	1276	1077	1077	1005	1005	645	645

Notes. The dependent variable is a dummy equal to one if Margaret is chosen. Treatment "Unbalanced" is a dummy equal to one if the participant is exposed to a pool of six male instructors, and zero if he/she is exposed to a gender balanced pool of instructors. In columns from 1 to 10, we report results of MTurk experiments for different samples of subjets, namely those who did not check any advice (1-2), those who checked at least one advice (3-4), those who checked at least two advices (5-6), those who checked at least 3 advices (7-8), and those who guessed corretly how many female instructors were in the pool (9-10). Robust standard errors are reported in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Summary Statistics of MTurk Experiment: Balanced versus Unbalanced (Male Scarce)

Group	Balanced			Unbalanced			(B-U)
	No.Obs	Mean	Std.Dev	No. Obs	Mean	Std.Dev	P-value
Panel A: Permutation variables							
Math Task	743	0.47	0.50	699	0.47	0.50	0.76
Margaret First	743	0.49	0.50	699	0.50	0.50	0.83
Margaret TA	743	0.49	0.50	699	0.50	0.50	0.80
Panel B: Sociodemographic variables							
Female	743	0.45	0.50	699	0.42	0.49	0.29
Age	743	35.78	11.32	699	34.67	10.32	0.051
White	743	0.77	0.42	699	0.75	0.44	0.28
College degree	743	0.60	0.49	699	0.70	0.46	0.00
Post-graduate degree	743	0.30	0.46	699	0.20	0.40	0.00
Panel C: Participants' performance							
Richard chosen	743	0.37	0.48	699	0.52	0.50	0.00
Duration	743	819.90	352.43	699	827.60	354.50	0.68
No. of advices	743	4.35	2.70	699	4.69	2.58	0.01
No. of correct answers	743	7.03	3.47	699	7.20	3.23	0.34

Notes. The group "Balanced" includes all participants exposed to a gender balanced pool of instructors, while the group "Unbalanced" includes all participants exposed to a pool of six female instructors. For each variable of interest, we report the number of observations, mean and standard deviation. The last column reports P-values of a t-test between variables in control and treatment group.

Table 5: Choice of Male Instructor when Male Instructors are Scarce

	(1)	(2)	(3)
Unbalanced (β)	0.136***	0.131***	0.161***
	(0.035)	(0.034)	(0.041)
Female×Unbalanced (γ)			-0.064
			(0.053)
Math Task	-0.005	-0.012	-0.012
	(0.028)	(0.028)	(0.028)
Margaret First	0.087**	0.089***	0.087***
	(0.033)	(0.033)	(0.032)
Margaret TA	-0.016	-0.013	-0.014
	(0.038)	(0.039)	(0.039)
Female		-0.020	0.012
		(0.028)	(0.036)
Age		-0.002	-0.002
		(0.001)	(0.001)
White		0.027	0.025
		(0.033)	(0.033)
College Degree		0.082*	0.085*
		(0.046)	(0.046)
Post-graduate Degree		0.026	0.028
		(0.046)	(0.046)
Constant	0.334***	0.338***	0.320***
	(0.044)	(0.069)	(0.070)
$\beta + \gamma$			0.096**
			(0.045)
R-squared	0.029	0.037	0.038
N	994	994	994

Notes. The dependent variable is a dummy equal to one if Richard is chosen. Unbalanced is a dummy equal to one if the participant is exposed to a pool of six female instructors, and zero if he/she is exposed to a gender balanced pool of instructors. All included participants checked the advice by the chosen instructor. Robust standard errors are reported in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Gender Gaps in Teaching Evaluations, by Field

Disciplines	Communication			Economics			Computer Science		
Column	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female_s	-0.154**			-0.273***			0.257*		
	(0.0626)			(0.0593)			(0.141)		
Female_p	0.00721	-0.0911		-0.0986	-0.0159		-0.322***	-0.335***	
	(0.166)	(0.167)		(0.165)	(0.171)		(0.0268)	(0.0389)	
$\text{Female}_S \times \text{Female}_P$	0.0160	0.0509	0.0548	0.283*	0.276**	0.290**	0.350	0.335	0.333
	(0.0989)	(0.0955)	(0.0859)	(0.152)	(0.118)	(0.119)	(0.286)	(0.275)	(0.274)
Constant	7.680***	7.544***	7.576***	9.217***	6.016***	6.482***	6.227***	5.764***	6.343***
	(0.484)	(0.354)	(0.317)	(0.515)	(0.531)	(0.680)	(0.809)	(0.429)	(0.399)
Course-Year FE	YES	YES	NO	YES	YES	NO	YES	YES	NO
Student FE	NO	YES	YES	NO	YES	YES	NO	YES	YES
Professor-Course-Year FE	NO	NO	YES	NO	NO	YES	NO	NO	YES
Student-Course Control	YES	YES	YES	YES	YES	YES	YES	YES	YES
Student Control	YES	NO	NO	YES	NO	NO	YES	NO	NO
R-squared	0.223	0.493	0.499	0.183	0.515	0.520	0.372	0.562	0.564
N	7,723	7,735	7,777	6,906	6,916	6,916	2,198	2,203	2,203

Notes. The dependent variable is the teaching evaluation score received by instructor i for course j. Evaluations in courses with less than six students are excluded from the analysis. Columns 1,4,7 include Course-Year fixed effects, Columns 2,5,8 include Course-Year fixed effects and Student fixed effects, and Columns 3,6,9 include Professor-Course-Year fixed effects and Student fixed effects. Standard errors, clustered at course-year level, are reported in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A Online Appendix

A.1 Description of MTurk Experiment

The experiment is structured in seven steps, which are listed below. In every step, participants are shown a screen window. In the first four steps, participants are free to choose when to move forward by clicking on the arrow in the lower right corner of the screen. Once the participants click on the arrow, they move to the next step and cannot go back. We made this rule clear by warning participants with this sentence at the bottom of the screen window in step 1 to step 4: "After a short while, you will be able to click on the arrow below in order to proceed. Once clicked, you will no longer be able to go back."

Step 1. All the participants are given the following information:[7]

- They will have to solve simple math/language tasks (10 questions) under time pressure.
- They will be paid based on performance (40 cents for each correct answer).
- They will all receive $1 for their participation.
- Before the test, they can read tips on how to solve the tasks written by different instructors.

Step 2. Two different lists of 6 instructors are shown to participants. They are not given any information other than the instructors' first names and qualification as "graduate student" (see Figure 1, panel A and B, upper part).

- Treatment participants are exposed to a pool of 6 male instructors.
- Control participants are exposed to a pool of 3 female and 3 male instructors.

Step 3. Participants are asked to choose one additional instructor; they can choose between one female and one male candidate (see Figure 1, panel A and B, lower part).

- The two candidates are Margaret (female candidate) and Richard (male candidate).
- The two candidates have the same educational background: they are both enrolled in a PhD.
- Participants are given some additional information about the two candidates: GPA and hours of experience as TA.

[7] Participants randomly assigned to the math task visualized precisely the following message: "Thank you for your participation in this study. You will receive 1 dollar for your participation, that is, if you complete the study. We estimate it will not take more than 15-20 minutes. We will ask you to perform a MATH task and we will pay you according to how well you do the task. In particular, we will ask you 10 questions with limited time to respond, and we will pay you 40 cents per correct answer. If you answer correctly all the 10 questions you will receive 4 dollars in addition to the 1 dollar for your participation. Before you do the task, you will be able to read explanations on the task, and you will receive tips on how to get the correct answer for the MATH questions quickly. You will have 10 seconds to answer each question. In the next screen you will find the pool of instructors, all of whom will explain the task and give you tips on how to solve the task correctly under limited time. After a short while, you will able to click on the arrow below in order to proceed. Once clicked, you will no longer be able to go back.". Participants randomly assigned to the English task visualized the same message, with the only difference that the word MATH was replaced by the word ENGLISH.

Step 4. Participants may read as many tips as they want. They do not have any time limit in this stage.

Step 5. Whenever they feel ready, participants can proceed with the exercise solving part. They have 10 seconds for each question.

- If participants are randomized into the math task, they have to solve 10 multiplications of the number 11 with a two or more digit number.
- If participants are randomized into the language task, they have to spell 10 English words correctly.

Step 6. Participants are asked to give some personal information (age, gender, education).

Step 7. At the end, participants are asked to answer the question "In the pool of six instructors how many women were there?". Options were in a range from zero to three.

Table A.1: Descriptive Statistics of Teaching Evaluations

	Comm.	Econ.	Comp. Sc.	Δ(E,CO) *P*-value	Δ(E,CS) *P*-value
Column	(1)	(2)	(3)	(4)	(5)
Panel A: Students Characteristics					
No. of Students	770	922	218	-	-
Dummy Female Student	0.69	0.43	0.12	0.00	0.00
Dummy Swiss Students	0.44	0.34	0.35	0.00	0.66
Dummy Italian Students	0.41	0.49	0.30	0.00	0.00
Dummy Other Nationalities	0.15	0.17	0.34	0.25	0.00
Dummy Bachelor Students	0.59	0.48	0.66	0.00	0.00
Student Age	24.56	23.89	24.41	0.00	0.03
Panel B: Course Characteristics					
No. of Courses	430	420	191	-	-
Dummy Compulsory Courses	0.60	0.45	0.71	0.00	0.00
Dummy Quantitative Courses	0.14	0.51	0.90	0.00	0.00
Class Size	34.30	39.36	24.61	0.05	0.00
Panel C: Instructor Characteristics					
No. of Instructors	181	171	89	-	-
Dummy Female Instructors	0.33	0.23	0.17	0.04	0.22
Dummy Full Professors	0.28	0.32	0.36	0.42	0.54
Dummy Associate Professors	0.13	0.17	0.18	0.33	0.84
Dummy Assistant Professors	0.07	0.10	0.10	0.26	0.96
Dummy Lecturers	0.52	0.40	0.34	0.02	0.30
Publish or Perish Citations	87.23	131.68	1225.48	0.11	0.00
Panel D: Student-Course Characteristics					
No. of Teaching evaluations (TE)	11,768	12,435	2,793	-	-
Dummy Students repeating courses	0.02	0.04	0.06	0.00	0.00
Dummy Students not reporting TE-Score	0.06	0.07	0.04	0.00	0.00
Student Grade	7.92	7.51	7.51	0.00	0.84
TE-Score: Overall satisfaction with the course	7.21	7.22	7.28	0.82	0.42

Notes. Table reports summary statistics related to students (Panel A), courses offered (Panel B), professors (Panel C), and students-course characteristics (Panel D) for the academic years 2015 to 2017. In each panel, we report sample numerosity in the first row. For each variable, we report the mean of the variable by faculty (Columns 1-3). In Column 4 we report the P-value of the difference between the mean values of Economics and Communication. In Column 5 we reports the P-value of the difference between the mean values of Economics and Computer Science.