Journal of Public Economics 193 (2021) 104323

Contents lists available at ScienceDirect

Journal of Public Economics

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

Reducing discrimination in the field: Evidence from an awareness raising intervention targeting gender biases in student evaluations of teaching

Anne Boring [a,b,*], Arnaud Philippe [c]

[a] Erasmus School of Economics, Netherlands
[b] LIEPP (Sciences Po), France
[c] University of Bristol, United Kingdom

ARTICLE INFO

Article history:
Received 7 March 2020
Revised 21 August 2020
Accepted 12 October 2020
Available online 28 December 2020

JEL classification:
C93
I23
J71

Keywords:
Student evaluations of teaching
Gender biases
Field experiment

ABSTRACT

This paper presents the results of a field experiment designed to reduce gender discrimination in student evaluations of teaching (SET). In the first intervention, students receive a normative statement reminding them that they should not discriminate in SETs. In the second intervention, the normative statement includes precise information about how other students (especially male students) have discriminated against female teachers in previous years. The purely normative statement has no significant impact on SET overall satisfaction scores, suggesting that a blanket awareness-raising campaign may be inefficient to reduce discrimination. However, the informational statement appears to significantly reduce gender discrimination. The effect we find mainly comes from a change in male students' evaluation of female teachers.

1. Introduction

Anti-bias awareness-raising campaigns are a common strategy that organizations use to reduce discrimination in the evaluations of employees and job candidates. These campaigns often convey a normative message highlighting how discrimination is wrong, unfair and should be avoided. Are such norm-setting strategies effective? Perhaps not, given that the biases driving discrimination are largely unconscious (Bertrand et al., 2005; Rooth, 2010; Oreopoulos, 2011; Glover et al., 2017). Individuals who discriminate may not believe that the normative message applies to their own behavior, because they may not be conscious of their own biases when they evaluate others. Including information to make biases conscious may therefore be a necessary condition for anti-bias awareness-raising campaigns to be effective.

This paper provides the result of a field experiment designed to test the impact of two types of awareness-raising campaigns—one with information to generate bias awareness, and one without—in the context of gender discrimination in student evaluations of teaching (SETs). This context resembles a common principle-agent problem, where an employer relies on information provided by a third party (such as managers, coworkers or clients) to evaluate an employee's performance and make personnel decisions. The principal would like to rely on unbiased information to make discrimination-free decisions. In the higher educational context, universities frequently rely on information from SETs for decisions involving the retention and promotion of instructors. However, studies from different countries provide empirical evidence that students can be biased in their evaluations of female instructors: in the U.S. (Arbuckle and Williams, 2003; MacNell et al., 2015), France (Boring et al., 2016; Boring, 2017), the Netherlands (Wagner et al., 2016; Mengel et al., 2019), Switzerland (Funk et al., 2019), and Australia (Fan et al., 2019) for instance. Universities worldwide therefore have a clear interest in reducing gender biases in SET scores to avoid making discriminatory personnel decisions.

We conducted the experiment in a French university where a study found evidence of gender biases in SET scores in previous years (Boring, 2017). The administration sent two different emails to students during the evaluation period. One email—the "purely normative" treatment—encouraged students to be careful not to discriminate in SETs. The other email—the "informational"

* Corresponding author at: Erasmus School of Economics, Netherlands.
E-mail addresses: boring@ese.eur.nl (A. Boring), arnaud.philippe@bristol.ac.uk (A. Philippe).

https://doi.org/10.1016/j.jpubeco.2020.104323
0047-2727/

treatment—added information to trigger bias consciousness. It included the same statement as the purely normative treatment, plus information from the study on gender biases in SETs. The message contained precise information on the presence of gender biases in SET scores in previous years at that university, including the fact that male students were particularly biased in favor of male teachers. The goal of this second treatment was to make the treated students explicitly aware of their own potential gender biases, by identifying with former students of the same university.

We created a difference-in-difference setting using the university's seven separate campuses. The students of two campuses were defined as controls: they did not receive any email during the three-week evaluation period. Three other campuses were treated with the normative message. The two remaining campuses were treated with the informational message. The administration sent the emails after some students had already completed their evaluations. This design provides us with a pre-treatment period for all campuses. Finally, the emails were sent to a random half of the students in each of the treatment campuses. This feature allows us to measure spillover effects of the treatments within campuses for the students who completed their SETs after the emails were sent. Campuses are located in different cities, which limits spillover effects between campuses: students communicate within campuses, but rarely across campuses.

Difference-in-difference analyses by teacher gender indicate that the purely normative treatment had no significant impact on reducing biases in SET scores. However, the informational treatment significantly reduced the gender gap in SET scores, by increasing the scores of female teachers. Overall satisfaction scores for female teachers increased by about 0.30 points (between 0.08 and 0.52 for the confidence interval at 5%), which represents around 30% of a standard error. The informational treatment did not have a significant impact on the scores of male teachers. These results are confirmed by a triple-difference analysis, in which we include all campuses and teachers. In all robustness checks, the informational treatment remains significant: when we compare campuses separately, when we look at men and women separately within each campus, and when we use the year before the experiment as control. Each strategy rests on slightly different hypotheses, but results remain consistent.

While the gender gap drops substantially, it does not close completely. In the campuses that received the informational treatment, the gender gap decreased by 0.09 points after the treatment, which represents a 47% decrease (from 0.19 to 0.10). This drop contrasts with an *increase* in the gender gap in the control group of 0.11 points, between before and after the moment the emails were sent on the treatment campuses. Over the same period in the year before, the gap had also increased (0.12 points) in the campuses where we ran the informational treatment. While these calculations are based on specifications using teacher fixed effects, they are indicative of the extent to which the bias can be reduced by such an intervention. The reduction in the gender gap following the informational email seems to be driven by male students increasing their scores for female teachers. On the informational treatment campuses, male students' mean ratings of female teachers increased from 2.89 to 3.20 after the emails were sent. This large increase in their scores to female teachers contributed to reducing the gap. There is no evidence that the informational email created (positive) discrimination by female students. Furthermore, the scores of the higher quality female teachers (those who generated more learning) seem to have been more positively impacted by the informational email. The effect of this treatment appears to have survived in the medium run, for the spring semester courses.

Finally, we find that the informational treatment had important spillover effects. On informational treatment campuses, we find an impact on students who received the email and on students who did not receive the email. Anecdotal evidence suggests that this email sparked conversations between students within campuses, de facto treating other students. We find weak empirical evidence of a small delay in the effect on students who did not directly receive the email. Information gathering (learning) following the informational treatment could therefore explain our results. We believe that these discussions probably contributed to making this treatment effective, whereas students in the other treated campuses are likely to have largely ignored the purely normative message. The persistence of the informational treatment effect in the medium run also appears to be more consistent with a learning explanation than with a purely behavioral explanation, and makes an experimenter demand mechanism unlikely. Furthermore, we test the effect of providing norms, with and without information, in a context where student behavior remains private information, therefore limiting the potential impact of a Hawthorne effect. Indeed, students complete their SETs online anonymously, thus excluding public scrutiny of their behavior as a possible mechanism to explain these results. However, since the informational email was longer and more precise than the normative email, we cannot fully rule out explanations based on saliency (the discrimination issue was more salient in the informational treatment) or priming (male students were specifically targeted in the informational treatment).

Our approach and setting have some limitations. In particular, the small number of treatment units limits statistical power. The mechanisms (learning, priming or saliency) are also difficult to disentangle. The results nonetheless convey important policy implications: while purely normative awareness-raising campaigns may be ineffective, including precise information on people's behavior can reduce discrimination. Such awareness-raising campaigns could have important spillover effects, but do not seem to create other forms of discrimination.

This paper contributes to the literature on the efficiency of interventions designed to reduce discrimination. Past research has shown that being *directly* informed of one's own biases through the use of implicit association tests seems to be an efficient strategy in the lab (Paluck and Green, 2009) and in the field (Alesina et al., 2018), although the evidence is still scant (Moss-Racusin et al. 2014; Bertrand and Duflo 2017). In our experiment, we study what happens when students are informed *indirectly* about their potential biases, using information from academic research. Importantly, we do not use a direct blaming and shaming approach, which the literature in other fields suggests may be counterproductive, for instance in firms' diversity trainings (Dobbin and Kalev, 2016).

Past research has focused on other strategies to reduce discrimination, through changes in the settings or rules in which firms make discriminatory decisions: organizing "blind auditions" (Goldin and Rouse, 2000), increasing the number of women in hiring committees (Kunze and Miller, 2017; Bagues, et al., 2017), using joint evaluations (Bohnet et al., 2015), and anti-discrimination laws (Collins, 2003, 2004). Our paper shows that providing information on people's behavior–a relatively easy strategy to implement–can be effective. In a recent lab experiment, Mengel (2020) tests an intervention similar to ours in the context of committee deliberations, and also finds that providing information on biases can be effective. These results suggest that real life interventions, such as informing committee members conducting interviews for European Research Council (ERC) grants about gender biases in recruitment committee decisions,[1] may effectively reduce discrimination.

[1] "Recruitment Bias in Research Institutes": https://www.youtube.com/watch?v=g978T58gELo.

We extend the results from Pope et al. (2018), who study the impact of a research-based awareness-raising campaign that arose following the publication of an NBER research paper by Price and Wolfers (2007) providing evidence of out-group bias in the fouls that referees call out in NBA games. Using a pre-post analysis, the authors find that the article's wide media attention caused a drop in discrimination in the following seasons (2007–2010). Our results shed light on mechanisms through which such awareness-raising campaigns can work in reducing biases in evaluations, by focusing exclusively on the change in behavior of the individuals who evaluate. Only students are treated in our context, whereas both those who evaluate (referees) and those who are being evaluated (players) are treated in the Pope et al. (2018) setting.

Finally, our results relate to the literature on the effectiveness of treatments aiming at inducing pro-social behavior through information (Asensio et al., 2014; Ida et al., 2013; Allcott and Rogers, 2014; Yoeli et al. 2013).

The paper is organized as follows. Section 2 describes the experiment. Section 3 presents the identification strategy and Section 4 the main results. Section 5 discusses the possible mechanisms. Section 6 concludes.

2. The experiment

2.1. Institutional setting

The field experiment took place in a selective French university specialized in social sciences (similar to a liberal arts school in the U.S.), in the fall semester of the 2015–16 academic year, on a cohort of 1570 students. Several features of the university's first year undergraduate studies are useful for the experiment. First, all first-year students must follow mandatory courses in history, political institutions, and microeconomics in the fall semester, and macroeconomics, political science, and sociology in the spring semester. Each course consists in two hours a week of a large lecture, plus two hours of classwork in small groups called seminars. The SET scores we analyze are from seminar courses, as there are many teachers, with enough variation in teacher gender (men teach most main lectures).

Second, we take advantage of the fact that undergraduate students are in seven separate campuses in different cities. Whereas each campus focuses on a different geopolitical area, students take the same mandatory courses in all campuses. At the end of their three years of study, all students receive the same degree in social sciences.

Third, the administration makes it mandatory for students to complete SETs online at the end of each semester. SETs take place over a three-week period, (the two final weeks of classes and the review week before final exams). By then, students have a good idea of what their continuous assessment grade will be, but not their overall course average (final grades are a weighted average of the continuous assessment and final exam grades). Furthermore, SETs remain anonymous to the teachers, who cannot trace back SET scores to individual students. Students who do not complete their SETs are unable to register for the following semester, thus guaranteeing a very high response rate. The same survey is used on all campuses (the questionnaire is in Appendix A).

Finally, a study on SETs in the Paris campus showed evidence of gender biases against female teachers in previous academic years (2008–2013) (Boring, 2017), with male students being particularly biased in favor of male teachers. Overall satisfaction scores were biased, as well as scores on different teaching dimensions. The dimensions that students valued in men and women tended to correspond to gender stereotypes. For example, women obtained higher scores in teaching dimensions such as course preparation and organization, while men scored higher on "contribution to intellectual development" and class leadership skills.

2.2. Treatments

The university's administration approved our experiment. We also received approval from J-Pal's Institutional Review Board (Appendix B).

The experiment consisted in sending two different emails to students. The normative treatment ("treatment one") encouraged students to avoid discrimination, especially gender discrimination (the full English version of the email in Appendix C). The email started with a generic statement about how evaluations are important to help the administration prepare courses for the following year. It then encouraged students to avoid discrimination, focusing more specifically on gender discrimination:

> *"Considering the importance of these evaluations, we would like to remind you that your evaluations must exclusively focus on the quality of the teaching and must not be influenced by criteria such as the instructor's gender, age or ethnicity. We ask you to pay close attention to these discrimination issues when completing your student evaluations. The goal is to avoid a situation in which, for instance, gender-based biases or stereotypes would systematically generate lower evaluations for women instructors compared to their male colleagues."*

This treatment resembles many anti-bias awareness-raising campaigns, whose main message is that "individuals should not discriminate". If biased individuals are not conscious that they discriminate, we hypothesize that this type of message is unlikely to be effective.

The informational treatment ("treatment two") added precise information to the normative statement. It explicitly stated that students had applied gender biases in the past, in the same context. By making treated students identify with students who were biased in the past, we hypothesize that this treatment may reveal to the treated students that they also might be biased. The second email (see Appendix C) drew students' attention to the working paper by Boring (2015), "*which suggests the existence of gender biases against female instructors of first year undergraduate seminars for all fundamental courses*". The email contained a link to the working paper, and presented its main results:

> *"the results of this study show that students tend to give lower ratings to their female instructors despite the fact that students perform equally well on final exams, whether their seminar instructor was a man or a woman. Male students in particular tend to rate male instructors higher in their student evaluations, although a slight bias by female students also exists. The differences in SET scores do not appear to be justified by other measures of teaching quality, such as an instructor's ability to make their students succeed on their final exams."*

The message included a graph showing that overall satisfaction scores were unrelated to student performance on the final exam, and that male students consistently gave higher overall satisfaction scores to male teachers. The email ended with the same normative statement as in treatment one.

2.3. Design

To measure the effects of the two treatments, we take advantage of the fact that the university has separate campuses. While students know each other quite well within each campus, they rarely communicate between campuses, enabling us to send

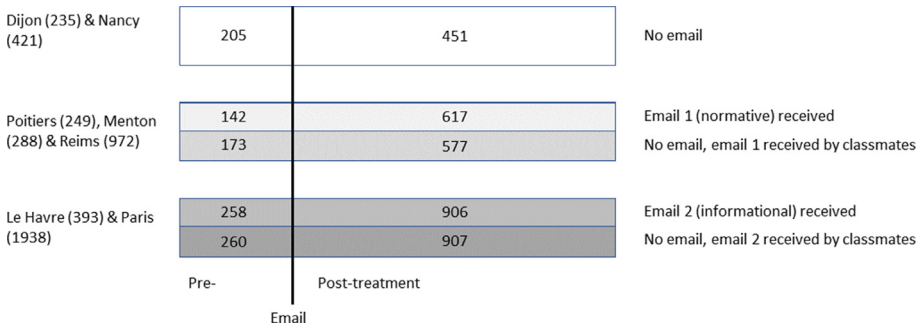


Fig. 1. Design of the experiment. Note: the numbers indicate the number of observations in the dataset (i.e. the number of evaluations). The black bar indicates the moment when the administration sent the emails. The campuses included in each group are indicated on the left-hand side.

different emails to students on different campuses. Fig. 1 presents the design of the experiment. The first treatment group includes students from three campuses: Menton (102 students), Poitiers (86 students) and Reims (337 students). We assigned the purely normative email to this group of students. The second treatment group includes students from the campuses in Le Havre (131 students) and Paris (657 students). We assigned the informational email to students from this second group. The other two campuses, Dijon (101 students) and Nancy (155 students), are the control group campuses. Each student completes three evaluations (one for each seminar). The dataset includes a total of 1509 evaluations for treatment one (95.8% response rate), 2331 evaluations for treatment two (98.5% response rate), and 656 evaluations for the control group (85.4% response rate).

We sent the emails to half the students on the treatment campuses. Before the beginning of the experiment, we randomly selected the students who would receive the emails. We use the following notations: group C is the control group; group TT1 (treatment treated one) includes all students who received the purely normative email; TC1 (treatment control one) includes all students who did not receive the email, but who were on the campuses that were treated with the purely normative email; TT2 and TC2 are similar to TT1 and TC1, but for the informational treatment campuses.

The university's gender equality officer sent the two emails simultaneously on a Friday evening, about one week after the beginning of the three-week evaluation period. Roughly one fifth of the evaluations had been completed by then: 20.9% in treatment one (normative) and 22.2% in treatment two (informational). In each treated campus, some evaluations were therefore completed before the treatment.

2.4. Data

Table 1 shows the descriptive statistics for the main student and teacher-related variables. 60% of the students are women. Almost all students are 18 years old. Students received higher continuous assessment grades (nearly 14 out of 20, on average), than final exam grades (11.7 out of 20, on average). Most students are French (73%). Finally, 32% of students were admitted through the international procedure, 10% of students were admitted through a specific procedure designed for students coming from lower income areas of France, and 46% were admitted through the main admissions procedure. The remaining students were admitted through a dual degree procedure.

A total of 155 teachers were evaluated during the fall semester: 20 in the control group, 39 in treatment one, and 96 in treatment two. Of these teachers, 39% are women (8 in the control group, 18 in treatment one, and 31 in treatment two). All but two teachers were evaluated both before and after the administration sent the emails. Most teachers obtained overall satisfaction scores that students qualified as "excellent" (39%) or "good" (40%). Only 6% received "insufficient", and 15% "average", overall satisfaction scores. Teachers generally teach on only one campus.

Table 1
Descriptive statistics on students and teachers.

	Mean	S.d.
Panel A. Students		
Share of women	0.60	0.49
Age	18.17	0.79
Continuous assessment (seminar) grade	139.86	22.46
Final exam grade	116.81	34.35
Share of students with French citizenship	0.73	0.44
Share of students admitted through specific procedure	0.10	0.31
Share of students admitted through entry exam (French high school)	0.46	0.50
Share of students admitted through international procedure	0.32	0.47
Share of students admitted through dual degree with a foreign university	0.08	0.27
Share of students admitted through dual degree with a French university	0.02	0.16
Share of students in a regular degree	0.79	0.41
Share of students in a dual degree with a foreign university	0.10	0.30
Share of students in a dual degree with a French university	0.11	0.31
Panel B. Teachers		
Share of women	0.39	0.49
Overall satisfaction scores	3.13	0.89
Share of "excellent" overall satisfaction scores	0.40	0.49
Share of "good" overall satisfaction scores	0.38	0.49
Share of "average" overall satisfaction scores	0.15	0.36
Share of "insufficient" overall satisfaction scores	0.06	0.24
History overall satisfaction scores	3.21	0.82
Microeconomics overall satisfaction scores	3.08	0.91
Political institutions overall satisfaction scores	3.09	0.93

3. Identification strategies

The experimental design includes features that enable us to use difference-in-difference and triple-difference analyses to measure the direct and indirect effects of the treatments. First, some campuses are treated while others are control. Second, evaluated teachers could be male or female. Third, on the treated campuses, some students had already completed their evaluations by the time the emails were sent, generating a pretreatment period. Fourth, only half of the students (random draw) received emails on the treatment campuses.

3.1. Difference-in-difference

In our first analysis, we eliminate any spillover effects by only including students from groups C, TT1 and TT2, i.e. the control

group and the groups in which students received emails. We exclude TC1 and TC2, i.e. the groups that could be affected by spillover effects.

Using groups C, TT1 and TT2, we run standard difference-in-difference regressions on female and male teachers separately. We use regressions of the form:

$$SET_{s,te,t} = \beta_0 + \beta_1 * TT1 + \beta_2 * TT2 + \beta_3 * post_t + \beta_4 * TT1 * post_t + \beta_5 * TT2 * post_t + \gamma * X_s + \delta_{te} + \varepsilon_{s,te,t} \quad (1)$$

where $SET_{s,te,t}$ is the evaluation of teacher te by student s at time t; $post_t$ is a dummy equal to one if t is after the mailing campaign; TT_1 and TT_2 are the two treatment groups; X_s are controls for student characteristics (gender, age, grades, nationality, admissions procedure); and δ_{te} are teacher fixed effects. Including teacher fixed effects overcomes the potential bias due to correlations between timing and teachers' characteristics.[2] Our variables of interest are β_4, which measures the effect of the normative treatment, and β_5, which measures the effect of the informational treatment.

Second, we measure the spillover effects of the treatments thanks to the two groups of students (TC1 and TC2) who did not receive an email, but who studied on the treatment campuses. We compare the SET scores of the students who belong to TC1 and TC2 after the mailing campaign, with the control, TT1 and TT2 groups. We run regressions of the form:

$$SET_{s,te,t} = \beta_0 + \beta_1 * TT1 + \beta_2 * TC1 + \beta_3 * TT2 + \beta_4 * TC2 + \beta_5 * post_t + \beta_6 * TT1 * post_t + \beta_7 * TT2 * post_t + \beta_8 * TC1 * post_t + \beta_9 * TC2 * post_t + \gamma * X_s + \delta_{te} + \varepsilon_{s,te,t} \quad (2)$$

where variables are similar to those in equation (1).

As in equation (1), β_6 and β_7 capture the effects of the emails on those who received them. In addition, β_8 and β_9 measure the spillover effects of the emails on TC1 and TC2. In equation (4) we are interested in the magnitude and statistical significance of β_8 and β_9, as well as in their differences with β_6 and β_7 (respectively). If β_8 and/or β_9 are equal to zero, then this would mean that the emails had no spillover effects. If β_8 (resp. β_9) is not statistically different from β_6 (resp. β_7), this would mean that the spillover effect was complete. We run equation (2) separately for female and male teachers.

Lastly, we measure the net effect of the treatments, i.e. the effect of the treatments on those who received email one or email two, and students around them. We run equation (1) with T1 and T2 instead of TT1 and TT2. This specification is especially interesting if the treatments had a very large spillover effect, and if TT1/TC1 and TT2/TC2 are very close.

3.2. Triple difference-in-difference

We measure the effect of the treatments in one single triple difference-in-difference. As the results are harder to read when using a triple difference-in-difference, we only use this strategy to measure the net effect of the treatment. We do so by running regressions of the form:

$$SET_{s,te,t} = \beta_0 + \beta_1 * Woman_{te} + \beta_2 * post_t + \beta_3 * T1 + \beta_4 * T2 + \beta_5 * post_t * Woman_{te} + \beta_6 * post_t * T1 + \beta_7 * post_t * T2 + \beta_8 * Woman_{te} * T1 + \beta_9 * Woman_{te} * T2 + \beta_{10} * post_t * T1 * Woman_{te} + \beta_{11} * post_t * T2 * Woman_{te} + \gamma * X_s + \delta * Z_{te} + \varepsilon_{s,te,t} \quad (3)$$

where variables are similar to those in equation (1).

In this equation, β_8 and β_9 capture the effect of the treatment on both male and female teachers. β_{10} and β_{11} capture the additional effect of the treatment on women in campuses of the treatment one and two (respectively).

Our identification strategy relies on the fact that the differences between students who completed their evaluations before and after the emails were sent are similar across groups. In Table 2 we test this hypothesis by running balancing checks on observable characteristics. We run our main regression – the triple difference-in-difference presented in Equation (3) – on observable characteristics instead of our main outcome (i.e. overall satisfaction scores). Table 2 confirms that our treatments are not correlated with observable characteristics. Out of eight regressions and 16 relevant coefficients – "post*female*T1" and "post*female*T2" – only two are marginally significant. Overall, students' characteristics do not seem to be correlated with the treatments.[3]

3.3. Alternative specifications

These strategies allow us to extract all the variations created by our design, and to compare the effect on male and female teachers. They are therefore our favorite specifications. We add two alternative strategies that can also shed light on our main results. First, we run difference-in-difference regressions in control, treatment one or treatment two campuses with female teachers as treatment, and male teachers as control. We regress teachers' SET scores on $post_t$, $post_t * Woman_{te}$ and teacher fixed effects for each campus type (control, treatment one, treatment two).

Second, we use students' evaluations from the previous year. This strategy enables us to run separate triple difference-in-difference analyses for each campus or each campus group (control, treatment one, treatment two). The three dimensions of the triple difference-in-differences are: a dummy equal to one if the teacher is a woman, a dummy equal to one if the evaluation is the year of the experiment, and a dummy equal to one if the evaluation is completed in the last two weeks of the evaluation period.

The main advantage of these specifications is that, by treating campuses or groups separately, the results could not be driven by cultural differences or sample size issues (the informational treatment group being larger than the two other groups). The main drawback is that they rely on the assumption that male teachers are a good control group.[4] The results of these alternative strategies are in the appendix.

4. Main effects

4.1. Graphical evidence

In Fig. 2, we present the evolution of the average overall satisfaction scores by teacher gender, groups (treatment two, treatment

[2] The results could be affected if the timing of the evaluation is correlated with students' characteristics. First, we control for observable characteristics in all regressions. Second, we include student fixed effects (in models presented in the appendix material). Including student fixed effects presents several limitations. It drastically reduces the power of the regressions by introducing numerous fixed effects. Also, because students mainly complete all their evaluations for the semester on the same day, we could only measure the effect if we used both the fall and the spring semester. The identification would come from the difference between the two semesters' scores among students who filled their evaluations for the fall semester before the treatments. For this reason, we do not use models with student fixed effects as our main specification, even though the results are similar.

[3] We do find, however, that students who complete their evaluations later tend to be lower performing students, in all campuses, as can be seen in the descriptive statistics by treatment group, and before versus after emails, in the appendix Table D1.

[4] Another reason why we do not present these alternative models in the core of the paper is that we had not mentioned them in the approval document we submitted to J-Pal's Institutional Review Board.

Table 2
Balancing checks.

	(1) Student female	(2) Final exam grade	(3) Cont. assessment grade	(4) French citizenship	(5) Age	(6) Entry specific procedure	(7) Entry exam	(8) International procedure
post	−0.0044	−3.83	−5.48***	−0.017	0.25***	−0.0047	−0.035	0.033
	(0.056)	(3.32)	(2.03)	(0.058)	(0.097)	(0.016)	(0.048)	(0.051)
post*T1	0.042	−1.42	2.89	−0.043	0.024	0.0056	−0.19***	0.081
	(0.085)	(5.27)	(3.77)	(0.085)	(0.15)	(0.026)	(0.063)	(0.065)
post*T2	−0.086	−2.01	0.48	−0.16**	−0.12	−0.0026	0.031	−0.010
	(0.068)	(4.14)	(2.81)	(0.070)	(0.11)	(0.028)	(0.055)	(0.052)
post*female	0.023	2.91	2.92	−0.056	−0.13	−0.031	0.058	−0.033
	(0.064)	(3.91)	(2.34)	(0.059)	(0.10)	(0.028)	(0.073)	(0.077)
post*female*T1	−0.025	3.83	−2.72	0.12	0.036	0.022	0.0071	0.0060
	(0.11)	(6.48)	(4.60)	(0.10)	(0.18)	(0.038)	(0.096)	(0.099)
post*female*T2	−0.029	−0.73	−6.06	0.033	−0.00074	0.090*	−0.14*	0.028
	(0.100)	(6.31)	(4.34)	(0.088)	(0.16)	(0.047)	(0.086)	(0.079)
Observations	4496	4473	4496	4496	4496	4496	4496	4496
Mean female teacher	0.59	115	140	0.73	18.2	0.096	0.48	0.32
Mean male teacher	0.61	118	140	0.76	18.1	0.11	0.48	0.30

Note: The dependent variable of each regression is specified in the column header. All regressions include teacher fixed effects. Coefficients of T1 and T2 are absorbed by the teacher fixed effects. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

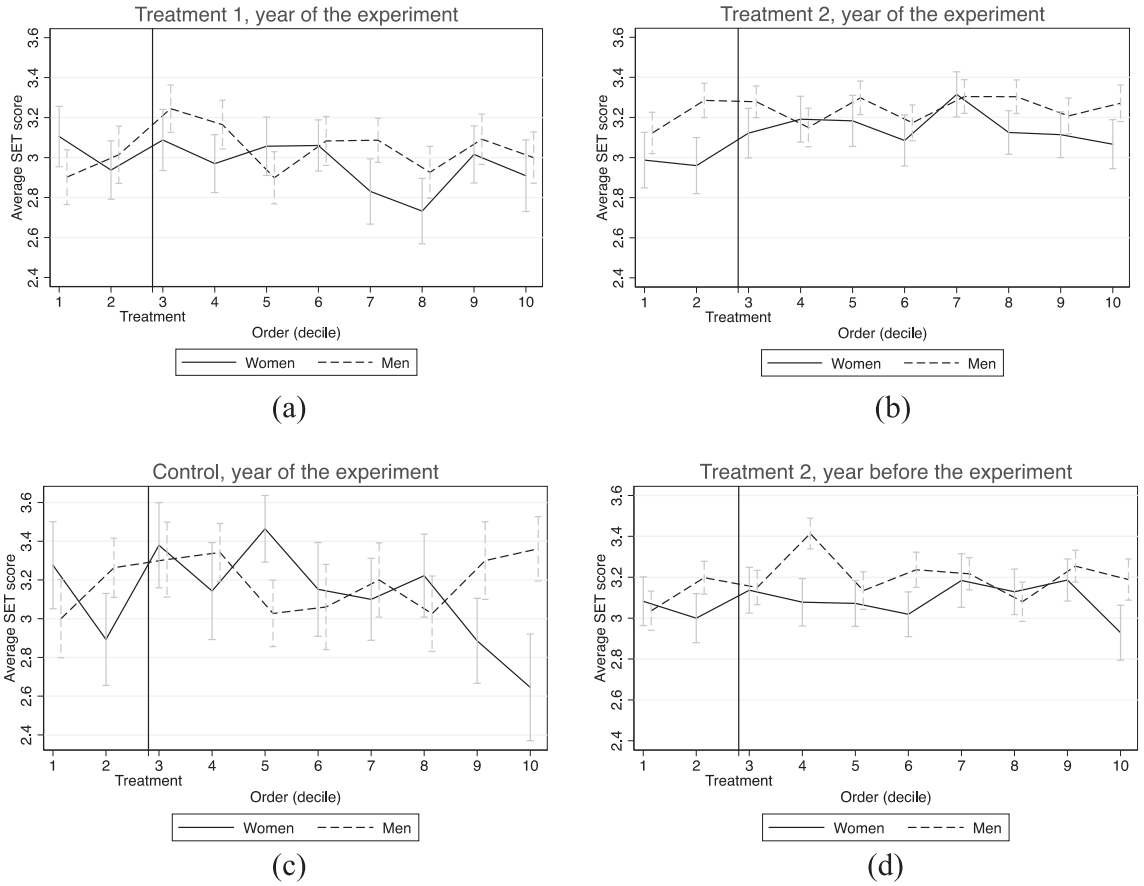


Fig. 2. Evolution of SET scores by teacher gender, and groups. Note: each point of each subgraph indicates the mean SET score by gender – women (solid line) or men (dashed line) – and period – first decile completed, second decile completed. . . Subgraphs (a), (b) and (c) present the evolution in treatment one, treatment two and control campuses respectively. Subgraph (d) presents the same evolution in treatment two campuses, the year before the experiment. Segments indicate the confidence interval at 10%. The vertical line corresponds to the moment when the administration sent-out the emails. The right-hand side of this bar is therefore the "post" period.

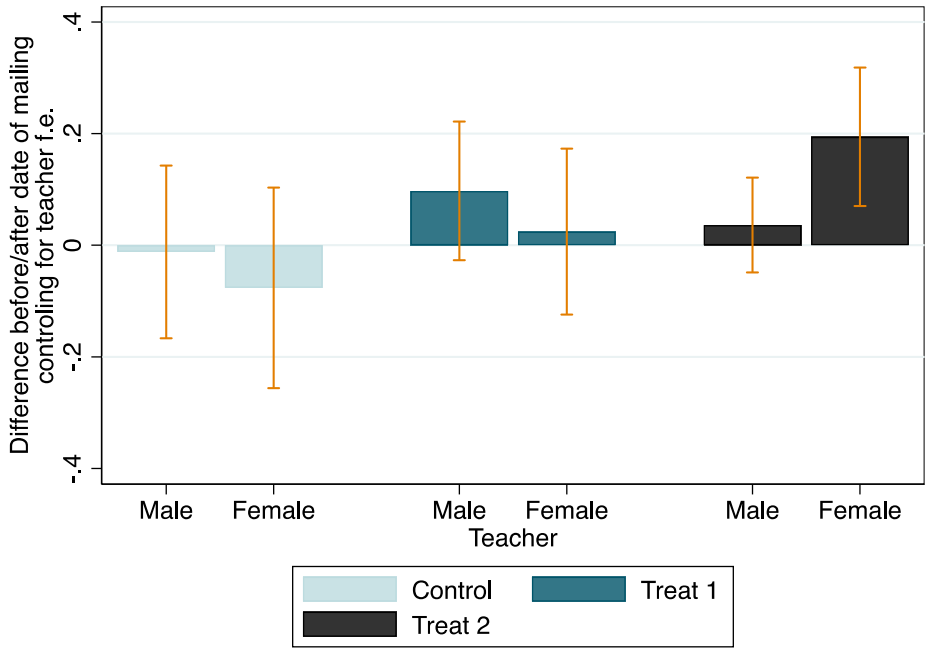


Fig. 3. Difference in SET scores before and after emails, after controlling for teacher fixed effects, by group and gender. Note: each bar indicates the difference in SET scores before and after emails for different groups (control, treatment one, or treatment two) and gender (women or men). Segments indicate the confidence interval at 5%.

one, control), over time. We split the SET scores for each gender and each group into ten subgroups based on when the students completed the evaluations: first decile (the first 10% that students completed in this group, for this teacher gender, during the evaluation period), second decile, etc. As the emails were sent after 20.9% (treatment one), and 22.2% (treatment two) were completed, the first two deciles constitute the pre-period, the third a "partially treated group" and the last seven deciles constitute the post period. Subfigure 2a presents the evolution in scores for male and female teachers in campuses included in treatment group one, 2b for treatment group two, and 2c for the control group. Subfigure 2d presents the evolution in campuses included in treatment group two for the year *preceding the experiment*.

On average, men's SET scores are greater than women's scores. In treatment group one and the control group, SET scores of both female and male teachers do not seem to have changed significantly after the emails were sent. Fig. 2a and c also do not show a clear gender gap before the treatment in these groups. However, Fig. 2b suggests that female teachers' scores in treatment two campuses increased after the treatment, whereas male teachers' scores do not seem to have been impacted by the treatment. The evolution observed in treatment two the year of the experiment did not occur the year before (subgraph 2d). Male students seem to mainly drive the change observed in treatment two campuses (appendix Fig. D1).[5] The graphical evidence suggests that treatment two increased women's SET scores. However, this pattern could be driven by the timing of the evaluations. Indeed, "good" and "bad" teachers could be evaluated at different points in time, and this evolution could drive the differences observed in Fig. 2. A simple way to tackle this issue is to measure the evolution of the SET scores by group and gender after controlling for teacher fixed effects. In this case, we simply aggregate the evolution by teacher. The results (Fig. 3) are consistent with the ones presented in Fig. 2: women's scores increased after the informational treatment.

4.2. Main results

Table 3 presents the main results of the effects of the two treatments on the overall satisfaction scores using difference-in-difference analyses (following Equation (1)). Regressions include controls for students' observable characteristics (age, whether the student is French, continuous assessment and final exam grades, average grades in other courses, and admission type), as well as teacher fixed effects.

The coefficients for the main variables of interest for women and men are shown in columns (1) and (2), where we restrict the dataset to the students who received the emails (TT1 and TT2) and the students of the control group. These regressions therefore exclude the control students who did not receive an email in the treatment campuses (TC1 and TC2). The results show that treatment two increased female teachers' SET scores (column 1). After the mailing campaign, the informational treatment induced a significant increase of 0.26 point for women. The purely normative treatment had no significant effect. The effects of treatment one and two are not statistically different. The effects of both treatments on male teachers' SET scores are not statistically significant.

In columns (3) and (4), we show the effect of the treatments in all groups following equation (2), as well as the p-values of the test of equality of the effects among subgroups. In these regressions, we therefore include all students, even those who did not receive an email in the treatment campuses (sample sizes are therefore larger than in columns 1 and 2). Once again, the results suggest that treatment two increased women's SET scores (column 3). This increase is observed both among those who received the email and those who did not, on the treatment two campuses. The difference between the effects on these two groups is not significant and the coefficients are similar (0.27 and 0.36 respectively). The spil-

[5] This finding is consistent with the fact that the gender gap in male students' scores was much larger (0.35 for male students, vs 0.08 for female students, in treatment two campuses for the pre-period, see Table D1). Furthermore, the informational email specifically mentioned that male students were particularly responsible for the gender gap.

Table 3
Main effects, fall semester courses.

	(1) Women	(2) Men	(3) Women	(4) Men	(5) Women	(6) Men	(7) All
Post	−0.079	0.016	−0.071	0.021	−0.072	0.021	0.026
	(0.090)	(0.078)	(0.090)	(0.076)	(0.089)	(0.076)	(0.077)
post*TC1			0.19	0.070			
			(0.14)	(0.11)			
post*TT1	0.091	0.17	0.078	0.17			
	(0.13)	(0.13)	(0.12)	(0.12)			
post*TC2			0.35***	0.017			
			(0.13)	(0.096)			
post*TT2	0.26**	0.054	0.26**	0.053			
	(0.13)	(0.099)	(0.13)	(0.098)			
post*T1					0.13	0.10	0.10
					(0.11)	(0.097)	(0.098)
post*T2					0.30***	0.035	0.032
					(0.11)	(0.087)	(0.087)
post*female							−0.11
							(0.12)
post*female*T1							0.024
							(0.15)
post*female*T2							0.28**
							(0.14)
Observations	1025	1542	1727	2746	1727	2746	4473
pval T1 T2	0.19	0.33			0.075	0.36	
pval TC1 TT1			0.40	0.40			
pval TC2 TT2			0.51	0.67			
pval TT1 TT2			0.13	0.30			
pval female*T1female*T2							0.03
Diff-in-diff	Yes	Yes	Yes	Yes	Yes	Yes	
Triple diff							Yes

Note: All regressions include teacher fixed effects and control variables for students (student gender, age, whether the student is French, and variables to control for academic ability and admissions type). Coefficients of variables TC1, TC2, TT1, TT2, T1 and T2 are absorbed by the teacher fixed effects in columns (1), (2), (5), (6) and (7). In order to simplify the table, coefficients of variables TT1 and TT2 are not presented in columns (3) and (4). They are not significant. Significance levels: *** p<0.01, ** p<0.05.

lover effect of treatment two seems to be complete. Men's scores did not change significantly following treatment two. Once again, we find no significant impact of treatment one on SET scores.

The fact that both the students who received the email and those who did not receive the email in treatment two campuses react similarly may be surprising. While our sample size does not enable us to further explore this result, anecdotal evidence indicates that students extensively discussed the email in treatment two campuses. In June, after the end of the year, we sent an email to students, asking whether they had discussed the content of the email with one another. Several students mentioned that they did indeed discuss the email with other fellow students.[6] The study was also published on the Facebook group of the feminist chapter on one campus, de facto treating other students from the campus. These anecdotal pieces of evidence are consistent with the timing of male and female teachers' SET evolution in treatment two among students who received or did not receive the email. While results are noisy because of small sample sizes, it seems that female teachers' SET scores by students receiving the email started to increase right after the email was sent. There appears to be a small lag in the increase in scores among those who did not receive the email, but who ended-up being treated through discussions with their peers (appendix Fig. D2).

Given this evidence of within campus spillover effects, we measure the effect of the treatments without distinguishing between students treated directly (those who received the email) and students treated indirectly (those who did not receive the email but who were in treated campuses). Results in columns (5) and (6) of Table 3 show that treatment two had a significant effect on women's SET scores, both in comparison to the control group (the coefficient is significant), as well as in comparison to the treatment one group (see the weakly significant p-value of the test of equality between the effect of treatments one and two, assuming complete spillover within each campus). Finally, this analysis confirms that treatment one does not appear to have had a statistically significant impact on either women or men.

These results are confirmed by triple-difference analyses. Column (7) shows the results of regressions including all overall satisfaction scores across all campuses. Female teachers in treatment two campuses received higher overall satisfaction scores after the emails were sent (the coefficient on $post_t * T_2 * Woman_{te}$ shows a statistically significant increase of 0.28 point, around 30% of the standard deviation of the outcome). The p-value of the test of equality between the effect of treatments one and two on female instructors is statistically significant. While the treatment does not close the gap in scores completely on the treatment two campuses, female teachers fare better following the emails. Their mean overall satisfaction scores increase from 3 to 3.15 out of 4 (Table D1). The gap in mean scores between female and male teachers shrinks, from a gap of 0.19 point before the email, to 0.10 after the email (about a 47% drop in the gap).

The effects presented in Table 3 are consistent with the finding that the purely normative statement had no effect on gender discrimination, while the informational treatment decreased it. We interpret these results as the effect of additional information on related behavior. This effect of providing information could be (partly or fully) driven by the discussion triggered by the email. Whether the informational treatment would have had the same effect in the absence of such discussions remains an open question.

Alternative explanations cannot be fully ruled out. Indeed, our results could also be interpreted as saliency–the issue of discrimi-

[6] For instance, one student said: "I remember this email very well because it created a long debate/discussion among my group of friends and I."

nation was more salient in the informational treatment– or priming– male students were specifically targeted. The fact that male students reacted while female students did not, makes an explanation based purely on saliency less likely. Our results could also be interpreted as an experimenter demand effect: students simply reacting in the expected direction. However, the fact that the effect survives in the middle run (see Section 5) provides some reassuring evidence that students did not only react when they were asked to do so. Lastly, the difference between the effects of treatment one and two could come from the original gender gap existing before the treatments. Indeed, as observed in Fig. 2, the gender gap in SET scores was large (−0.19 for women) and significant in treatment two campuses but not significant in control and treatment one campuses. In this context we could not exclude that treatment two worked better simply because there was more room for change.

In the appendix Table D2, we show that our main results are robust to perturbations of the main specification. In the first three columns we present the results of the triple difference-in-difference when using ordered logits instead of ordinary least squares, controlling for student fixed effects, and controlling for both teacher and student fixed effects (in a model including SET scores of both the fall and spring semesters). The results remain similar in magnitude and significant at the 5% level. The next three columns deal with various ways of clustering the standard errors. We first cluster at the campus level, as this is the main unit of randomization (column 4).[7] We then cluster at the student (column 5) and teacher (column 6) levels, as exogeneous shocks could affect all the SETs completed by one student or all the evaluations of one teacher. While these various ways of clustering seem legitimate, Table D2 indicates that the significativity of the main result is not affected by the clustering choice. Using a binary outcome, we find that all the effect comes from the margin between "average" and "good" overall satisfaction scores. No effect is observed at the margins between "bad" and "average" or between "good" and "excellent". The fact that the effect is localized at one margin and not from a shift of the entire distribution makes an explanation based on a behavioral response less likely.

Our results are also robust to alternative specifications. First, using difference-in-differences in each group with male teachers as control leads to similar results (Table D3). In comparison to men, women tend to obtain higher SET scores after the emails in treatment two (column 3), but not in control (column 1) nor treatment one (column 2) campuses. Second, using SET scores from 2014-2015 (the preceding year), we measure, for each campus (Table D4, columns 1–7) or group (Table D4, columns 8–10), the effects of completing evaluations in the last two weeks of the evaluation period,[8] studying in 2015–2016 (the year of the experiment), and the interaction of the two: completing the evaluation in the post period when studying in 2015–2016 (the period of the treatment). We find results similar to the ones presented in Table 3: no effect of the purely normative treatment, and an increase in female teachers' SET scores after the informational treatment. We also find that the effect of the informational treatment is observed in both Paris and Le Havre, the two treated campuses, even though the email presented results obtained in Paris exclusively. These results rule-out the idea that Parisian students were particularly primed by the informational email. This analysis also suggests that cultural differences of students in treatment one versus treatment two campuses are unlikely to be driving the results. Indeed, treatment one had no significant effect in any of the three treatment one campuses, whereas we find an effect in both treatment two campuses.

We also find no effect in a placebo exercise (appendix Table D2, column 10).

5. Mechanisms

We first focus on the differences of the effects based on student gender. Indeed, Boring (2017) found that male students were the ones who had a bias in favor of male teachers, generating higher overall satisfaction scores for male teachers in the Paris campus. The treatment two email explicitly referred to this difference among students. Two mechanisms could drive our main results. First, male students, who were mainly responsible for the gender gap in scores, may have corrected their biases following the information they received. Second, female students may have tried to counterbalance the biases through positive discrimination.

In order to further investigate these hypotheses, we run our main model on male and female students separately. Results are presented in the first two columns of Table 4. They show that, after treatment two, male students gave higher overall satisfaction scores to female teachers. Female students were not affected. Even though the difference between the two effects is not statistically significant because of small sample sizes, these results indicate that the informational treatment seems to have reduced male students' gender biases, without creating positive discrimination. Male students' mean ratings of female teachers increased by 0.31 points, from 2.89 to 3.2 after the informational email. Although there remains a 0.12-point gap with the way that male students rated male teachers on treatment two campuses (mean score of 3.32 for male teachers after the email), this gap is smaller than the one in the *post* period in the control group (0.21) and in treatment one (0.26). As a final comparison, the gap found by Boring (2017) on the Paris campus in previous years, between the way that male students rated male and female teachers' overall satisfaction scores, was equal to 0.19.

Second, we measure whether treatment two impacted all female teachers or mainly benefited the higher quality teachers. We define a "good teacher" as a teacher who generated more learning in students, measured as a teacher whose students received higher average grades on the final exam (above the median grade within campus). Teaching to the test and grade leniency are not an issue for the final exam in our context. Indeed, the final exam is designed by the main lecturer and is common to all students. This feature enables us to compare student learning across instructors. Students' final exams are also graded by a different teacher than the one they had during the semester. The grading is anonymous (double blind), and it takes place after students have completed their SETs for the semester. This feature specifically prevents grade leniency. Results of regressions separating the better teachers from the other teachers are presented in columns (3) and (4) of Table 4. They indicate that the higher quality female teachers especially benefitted from the higher overall satisfaction scores with treatment two.

Third, we measure whether "good" students reacted differently. We define "good" students as those who obtained above the median final grades within campus. Results are presented in columns (5) and (6) of Table 4. This analysis does not yield statistically significant results, suggesting that both types of students may have increased the overall satisfaction scores of female teachers.[9]

[7] While natural, this level of clustering is problematic because of the small number of units (only 7). For this reason, we did not use it in the main specification.

[8] More precisely, we divide the evaluation period in two: the first 23% and the remaining 77%.

[9] Results are similar when "good students" are defined as students who get final grades above the median within campus *in other courses* (not shown).

Table 4
Mechanism of the effect.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Students		Teacher's quality		Student's level		Length of the effect
	Male	Female	>median	<median	>median	<median	
post	0.021	0.015	0.028	0.030	0.067	−0.033	0.030
	(0.11)	(0.11)	(0.12)	(0.100)	(0.12)	(0.099)	(0.078)
post*T1	0.24	0.077	0.19	0.076	0.013	0.21	0.076
	(0.16)	(0.13)	(0.17)	(0.12)	(0.14)	(0.13)	(0.099)
post*T2	0.037	0.047	−0.0016	0.066	−0.030	0.081	0.021
	(0.13)	(0.12)	(0.13)	(0.12)	(0.13)	(0.12)	(0.088)
post*female	−0.16	−0.065	−0.14	−0.066	−0.011	−0.24	−0.11
	(0.18)	(0.16)	(0.15)	(0.23)	(0.16)	(0.17)	(0.12)
post*female*T1	0.17	−0.091	0.083	−0.11	−0.11	0.17	0.059
	(0.24)	(0.20)	(0.21)	(0.26)	(0.21)	(0.22)	(0.15)
post*female*T2	0.44**	0.14	0.32*	0.21	0.21	0.33	0.27*
	(0.22)	(0.19)	(0.18)	(0.25)	(0.19)	(0.21)	(0.14)
Spring semester							0.024
							(0.12)
Spring*T1							0.14
							(0.15)
Spring*T2							0.078
							(0.14)
Spring*female							−0.27*
							(0.15)
Spring*female*T1							0.15
							(0.20)
Spring*female*T2							0.56***
							(0.20)
Observations	1766	2707	2154	2319	2369	2104	8655

Note: Fall semester only in the first six columns. Fall and spring semesters in column (7). All regressions include control variables and teacher fixed effects. Coefficients of variables T1 and T2 are absorbed by the teacher fixed effects in all columns. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Fourth, we measure whether the treatments had any medium run effect. We do so by introducing the spring semester SET scores in the sample, and running our main regression with additional parameters for "spring"; "spring*T1"; "spring*T2"; spring*female"; spring*female*T1"; and "spring*female*T2". The effect of the informational treatment remains significant during the spring semester: female teachers improved their scores. The normative treatment remained ineffective. The pre period becomes the reference for the evaluations filled both in the fall semester after the emails were sent (for similar courses given by same teachers), and in the spring semester (with other courses and mostly other teachers). The identification of the effect in the spring semester is therefore weaker. Nonetheless, these results do not fit with an explanation based on a purely behavioral response to the email. As students' behavior changed in the medium run, this result seems to indicate that students gained a better understanding of their own behavior.

In appendix Table D5, we explore the effect of the treatments on the different teaching dimensions that the students also evaluate. Surprisingly, while only treatment two decreased the gender gap in overall satisfaction scores, the two different treatments seem to have the same effect on the teaching dimensions and may have actually reinforced gender stereotypes.[10] For instance, women's scores in "quality of instructional materials" or "clarity of course assessment" are significantly better after both treatments. These questions tended to be more favorable to female instructors, according to Boring (2017). We also find that all teachers' scores in "contribution to intellectual development" are significantly higher after the treatment. Boring (2017) found that (male) students were particularly biased on this item, as well as on an instructor's class leadership/quality of animation skills. We do not find evidence of a closing of the gender gap for these two questions. Both items were mentioned in the informational treatment, but it seems that students did not pay as much attention to these items, possibly because the treatment mentioned them only briefly and towards the end of the email. Other teaching dimensions do not seem to be impacted. However, we do find a positive effect of treatment two for women on what should be an objective item: how many grades students received during the semester. This result confirms previous research suggesting that students' evaluations remain subjective even on what should be purely objective criteria (Boring et al., 2016).

6. Conclusion

What constitutes an effective way to educate students about their own biases, and how can universities apply these results to reduce biases in SETs in their own contexts? One policy advice would be to remain cautious about the content of the awareness-raising message the administration sends to students. Our results suggest that simply telling students not to discriminate using a blanket administrative statement is likely to be ineffective. However engaging students in discussions about the role that discrimination plays in SET scores and presenting them with the large body of evidence that now exists can be efficient to reduce discrimination in scores. This strategy can be useful for universities that are reconsidering their use of SETs because of the existence of gender biases.

How should universities engage students about discrimination in SET scores? Some instructors may worry that, if they are the ones who try to encourage students to treat all professors equally, the intervention may backfire against them individually. To avoid uncomfortable and potentially counterproductive situations for instructors, it may be neces-

[10] When we adjust the p-values for multiple hypothesis testing in this Table by calculating Anderson's sharpened False Discovery Rate q-values, we find that the effects are not significant, suggesting that these results may not be very robust.

sary for the intervention to be carried-out by the administration, and not the instructors being evaluated. Furthermore, the administration must beware to avoid potential counterproductive activations of stereotypes through the anti-bias intervention (Dobbin and Kalev, 2018).

Further work needs to be conducted in order to fully understand the type of information that maximizes the impact. In the context of this research we used precise information on discrimination within the same university. The efficiency of similar messages displaying results obtained in another university or on related but not identical behavior remains an open question.

Finally, we believe that our results have broader implications. One of the main conclusions of our field experiment is that the content of an awareness-raising campaign is important. Poorly designed messages can be ineffective, which may explain the persistence of discrimination despite millions of dollars spent every year by firms, governmental agencies and non-governmental organizations on anti-discrimination campaigns. Our results suggest that these campaigns, which resemble our normative treatment, are likely to be inefficient. Similar results have been found on the efficiency of awareness-raising health campaigns, such as information campaigns designed to reduce anti-vaccination beliefs (Horne et al., 2015; Nyhan et al., 2014; Nyhan and Reifler, 2015).

Acknowledgements

The authors are grateful for the helpful comments and suggestions by Michèle Belot, Robert Dur, Mathilde Guergoat-Larivière, Nagore Iriberri, participants of the AFSEE, EALE, Advances with Field Experiments conferences, Milan Labour Lunch Seminar, and 2018 Field Days conference, as well as Bank of Spain, CREST, IAST, University of Bristol, Erasmus School of Economics, International Institute of Social Studies, George Mason University, Paris Business School, National University of Singapore and Paris School of Economics seminar participants. The authors would also like to thank the anonymous referees who provided very helpful suggestions, as well as Stéphane Auzanneau for his crucial help collecting the data. This project has received funding from the European Union's Seventh Framework Program for research, technological development and demonstration under grant agreement no 612413. Support through the ANR Labex IAST is gratefully acknowledged.

Appendix A. The survey questions

Students can give an "excellent", "good", "average", "insufficient" or "not applicable" rating to the following questions:

(1) How do you evaluate the preparation and the organization of classes?
(2) How do you evaluate the quality of the teaching materials?
(3) How do you evaluate the clarity of the assessment criteria?
(4) How do you evaluate the usefulness of feedback?
(5) How do you evaluate your teacher's class leadership / quality of animation skills?
(6) How do you evaluate your teacher's ability to encourage group work?
(7) How do you evaluate your teacher's availability and communication skills?
(8) How do you evaluate the course's ability to relate to current issues?
(9) How do you evaluate your teacher's contribution to your intellectual development?
(10) **What is your overall level of satisfaction?**

The survey includes the following additional questions:

(11) Compared with other courses this semester, I invested much more effort / as much effort / much less effort in this course.
(12) How many assessments did you have throughout the semester? 0 to 2, 3 to 4, 5 to 6, 7 or more.
(13) Were written assignments given back within deadlines? Yes / No
(14) Were oral presentation grades given back within deadlines? Yes / No

Finally, the survey includes two open-ended questions (which we do not analyze here):

(15) What are the strong points of this course?
(16) What are the points that the teacher could improve?

Appendix B. Approval by the IRB

ABDUL LATIF JAMEEL Poverty Action Lab J-PAL EUROPE	**Dossier n°**	IN/2015-008
	Date	18 12 2015

Décision de l'IRB de J-PAL Europe

Chercheurs principaux : Anne BORING, Arnaud PHILIPPE

Intitulé de l'étude : Diminuer les biais de genre : expérience randomisée sur les évaluations des enseignements

Demande initiale

Date de la décision : 18 décembre 2015

Date d'expiration: 17 décembre 2016

[x] Approuvé

Cette étude ne présente pas de risque pour les sujets humains. Les connaissances qui résulteront de cette étude sont suffisantes pour justifier sa mise en œuvre.

J-PAL EUROPE
PSE-Ecole d'économie de Paris
AP-HP
1 place du Parvis Notre Dame
75004 Paris
+33(0)1 43 29 70 81

IRB J-PAL Europe – Fiche décision

Appendix C. The two emails

Each email included a message in French, followed by the English version. We only reproduce the English versions below.

Email one (purely normative treatment):

Dear Student,

This fall semester's student evaluations of teaching are open since Monday November 23rd. These evaluations, which are mandatory for students to complete, are read by your instructors and closely analyzed by the *Direction des études et de la scolarité* in order to prepare the upcoming academic year. Your comments are extremely useful for the administration of Sciences Po in order to improve the quality of our programs, in close collaboration with our teaching staff.

Considering the importance of these evaluations, we would like to remind you that your evaluations must exclusively focus on the quality of the teaching and must not be influenced by criteria such as the instructor's gender, age or ethnicity. We ask you to pay close attention to these discrimination issues when completing your student evaluations. The goal is to avoid a situation in which, for instance, gender-based biases or stereotypes would systematically generate lower evaluations for women instructors compared to their male colleagues.

Best regards,

Signature of the Gender Equality Officer

Email two (informational treatment):

Dear Student,

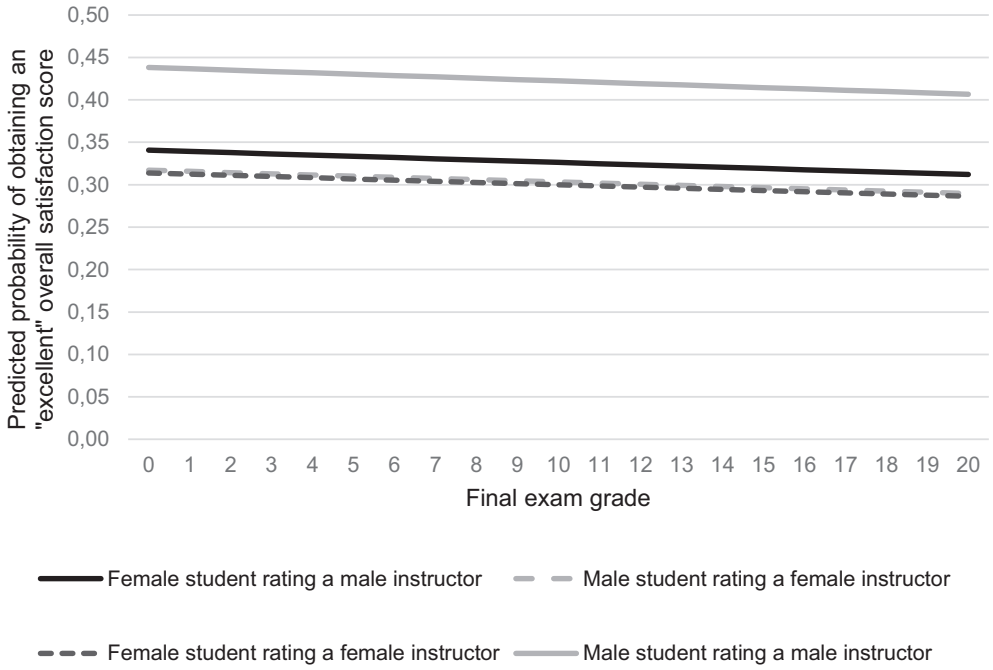
In this period of student evaluations of teaching (SET), we would like to bring your attention to the results of a recent study which suggests the existence of gender biases against female instructors of first year undergraduate seminars (i.e. the *conférences de méthode*) for all fundamental courses.

Indeed, the results of this study show that students tend to give lower ratings to their female instructors despite the fact that students perform equally well on final exams, whether their seminar instructor was a man or a woman. Male students in particular tend to rate male instructors higher in their student evaluations, although a slight bias by female students also exists. The differences in SET scores do not appear to be justified by other measures of teaching quality, such as an instructor's ability to make their students succeed on their final exams.

Let's take the example of students whose seminar average grade is 13.5 and the final exam grade is 12 (these grades correspond to the student averages observed during the period 2008–2013, pooling all fundamental courses together). Given these students, female seminar instructors have a 30% chance of obtaining an "excellent" overall satisfaction score, from both male and female students (and keeping constant course characteristics, such as the day and time of class). Given these grades, however, male instructors have a 33% of obtaining an "excellent" overall satisfaction score when evaluated by a female student and even a 42% chance when evaluated by a male student. These results mean that given an equal performance on exams, female instructors are 19% less likely to obtain "excellent" overall satisfaction scores compared to male instructors (taking into account the proportion of male and female students). These differences are statistically significant.

Furthermore, male students systematically rate male instructors higher, no matter students' results on final exams, as shown in the graph below.

Graph: Correlation between students' final exam grades and the predicted probability of giving an "excellent" overall satisfaction score, by student and instructor gender



Finally, the results of this study suggest that students apply gender stereotypes in the way they respond to more specific questions, such as an instructor's class leadership/quality of animation skills or the ability to contribute to students' intellectual development.

Given these results, we would like to remind you that your evaluations must exclusively focus on the quality of the teaching and must not be influenced by criteria such as the instructor's gender, age or ethnicity. We ask you to pay close attention to these discrimination issues when completing your student evaluations. The goal is to avoid a situation in which, for instance, gender-based biases or stereotypes would systematically generate lower evaluations for women instructors compared to their male colleagues.

Best regards,

Signature of the Gender Equality Officer

Appendix D. Additional material

This appendix presents additional tables mentioned in the paper.

Table D1 shows descriptive statistics by group (control, treatment one, and treatment two) and by time period (before and after the emails).

Tables D2, D3 and D4 present some robustness checks of our main results. In Table D2 we present perturbations of our main specification. Column (1) presents the results when using ordered

Table D1
Descriptive statistics on students and teachers, by campus group, before and after emails.

	Control				Treatment one				Treatment two			
	Before		After		Before		After		Before		After	
	Mean	S.d.	Mean	S.d.	Mean	S.d.	Mean	S.d.	Mean	S.d.	Mean	S.d.
Panel A. Students												
Share of women	0.58	0.49	0.60	0.49	0.69	0.46	0.58	0.49	0.60	0.49	0.61	0.49
Age	18.24	0.81	18.49	1.04	18.09	0.60	18.24	0.83	17.88	0.43	18.04	0.63
Continuous assessment (seminar) grade	147.29	24.37	143.9	24.81	140.79	21.36	139.91	22.95	141.6	21.27	137.32	21.45
Final exam grade	118.36	33.55	114.86	35.86	125.74	30.19	122.54	32.27	113.79	34.61	112.64	35.2
Share of French citizenship	0.58	0.49	0.55	0.50	0.75	0.44	0.58	0.49	0.97	0.18	0.86	0.35
Share of Specific admissions procedure	0.03	0.17	0.01	0.11	0.04	0.20	0.04	0.19	0.18	0.38	0.18	0.38
Share of Entry exam admissions procedure	0.25	0.44	0.25	0.43	0.49	0.50	0.29	0.45	0.68	0.47	0.63	0.48
Share of International admissions procedure	0.70	0.46	0.71	0.45	0.38	0.49	0.51	0.50	0.05	0.22	0.09	0.29
Panel B. Teachers												
Share of women	0.44	0.50	0.46	0.50	0.46	0.50	0.39	0.49	0.33	0.47	0.36	0.48
Share of "Excellent" scores	0.44	0.50	0.42	0.49	0.36	0.48	0.35	0.48	0.42	0.49	0.44	0.50
Share of "Good" scores	0.37	0.48	0.36	0.48	0.38	0.49	0.40	0.49	0.36	0.48	0.38	0.49
Share of "Average" scores	0.13	0.33	0.14	0.35	0.15	0.36	0.17	0.38	0.15	0.36	0.14	0.34
Share of "Insufficient" scores	0.06	0.24	0.07	0.26	0.11	0.31	0.08	0.27	0.07	0.25	0.04	0.20
History scores	3.00	0.89	3.05	0.84	3.07	0.88	3.09	0.87	3.20	0.86	3.39	0.72
Law scores	3.32	0.89	3.21	0.96	2.84	1.03	3.03	0.95	3.10	0.94	3.11	0.88
Microeconomics scores	3.24	0.84	3.12	0.93	3.04	1.01	2.91	0.91	3.10	0.93	3.15	0.87
Scores: male students for female teachers	3.28	0.93	3.07	0.98	2.96	1.04	2.96	0.92	2.89	0.98	3.20	0.83
Scores: male students for male teachers	3.37	0.71	3.28	0.81	2.85	1.10	3.22	0.90	3.24	0.93	3.32	0.82
Scores: female students for female teachers	3.14	0.93	3.08	1.00	3.13	0.87	2.92	0.91	3.08	0.88	3.10	0.86
Scores: male students for male teachers	3.09	0.90	3.12	0.87	2.94	0.98	2.96	0.90	3.16	0.88	3.21	0.83

Note: "Scores" refer to overall satisfaction scores.

Table D2
Robustness checks.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Ordered logit	Fall and spring semester	Fall and spring semester	Fall semester, standard errors clustered by:			Dummy			Placebo: year before the experiment
		Student fixed effects	Teacher and student fixed effects	campus	teacher	student	1 vs 2/3/4	1/2 vs 3/4	1/2/3 vs 4	
Post	0.012	-0.048	-0.067	0.026	0.026	0.026	0.011	0.0087	0.0058	-0.016
	(0.23)	(0.12)	(0.090)	(0.068)	(0.072)	(0.080)	(0.021)	(0.036)	(0.051)	(0.086)
Post*T1	0.34	-0.034	0.14	0.10	0.10	0.10	0.047	0.042	0.016	0.10
	(0.28)	(0.15)	(0.11)	(0.078)	(0.10)	(0.10)	(0.029)	(0.046)	(0.063)	(0.11)
Post*T2	0.094	0.029	0.020	0.032	0.032	0.032	0.012	0.014	0.0061	0.17*
	(0.26)	(0.14)	(0.10)	(0.069)	(0.084)	(0.092)	(0.025)	(0.041)	(0.057)	(0.094)
Post*female	-0.26	-0.17	-0.065	-0.11***	-0.11	-0.11	-0.029	-0.048	-0.030	0.011
	(0.37)	(0.14)	(0.11)	(0.020)	(0.085)	(0.12)	(0.039)	(0.053)	(0.074)	(0.13)
Post*female*T1	0.027	-0.058	0.0035	0.024	0.024	0.024	-0.010	0.0053	0.030	-0.11
	(0.45)	(0.19)	(0.15)	(0.099)	(0.13)	(0.15)	(0.050)	(0.071)	(0.092)	(0.16)
Post*female*T2	0.82*	0.36**	0.29**	0.28**	0.28**	0.28*	0.040	0.14**	0.10	-0.21
	(0.43)	(0.17)	(0.14)	(0.076)	(0.13)	(0.14)	(0.046)	(0.065)	(0.087)	(0.15)
Observations	4,473	8,655	8,630	4,473	4,473	4,473	4,473	4,473	4,473	4,398

Note: all regressions include teacher fixed effects and control variables for students (student gender, age, whether the student is French, and variables to control for academic ability and admissions type). Coefficients of variables T1 and T2 are absorbed by the teacher fixed effects. Columns (1) and (4) to (9) contain all the evaluations completed in the fall semester 2015-2016. Columns (2) and (3) contain evaluations from the fall and the spring semesters 2015-2016. Column (10) contains all the evaluations completed in the fall semester 2014-2015. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table D3
Robustness checks, difference-in-difference with male as control group.

	(1)	(2)	(3)
	Difference in difference		
	Control	Treatment 1	Treatment 2
Post	0.022	0.11*	0.048
	(0.074)	(0.061)	(0.043)
Post*female	−0.078	−0.079	0.18**
	(0.12)	(0.094)	(0.077)
Observations	654	1503	2316

Note: All regressions include teacher fixed effects and control variables for students (student gender, age, whether the student is French, and variables to control for academic ability and admissions type). Significance levels: ** p<0.05, * p<0.1.

Table D4
Robustness checks, using year 2014–2015.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Triple difference in difference using the preceding year									
	Dijon	Nancy	Menton	Poitiers	Reims	Le Havre	Paris	Control group	T1	T2
Post	0.11	0.026	0.039	−0.10	0.19*	0.13	0.15***	0.042	0.081	0.15***
	(0.12)	(0.11)	(0.15)	(0.12)	(0.11)	(0.13)	(0.040)	(0.085)	(0.072)	(0.038)
Post*female	0.029	−0.038	−0.13	0.24	−0.23	−0.41**	−0.16**	−0.0072	−0.10	−0.21***
	(0.21)	(0.16)	(0.21)	(0.18)	(0.14)	(0.20)	(0.069)	(0.13)	(0.098)	(0.066)
Post*year2015	−0.31*	0.049	−0.068	0.32*	−0.061	−0.20	−0.085	−0.015	0.039	−0.10*
	(0.17)	(0.14)	(0.24)	(0.19)	(0.13)	(0.17)	(0.061)	(0.11)	(0.092)	(0.057)
Post* year2015	0.023	−0.044	0.54	−0.22	0.051	0.93***	0.27**	−0.076	0.010	0.39***
*female	(0.28)	(0.21)	(0.54)	(0.25)	(0.17)	(0.29)	(0.11)	(0.17)	(0.13)	(0.10)
Observations	475	814	544	507	1543	720	4268	1289	2594	4988

Note: The sample includes the evaluation completed in the first semester of school year 2014–2015 (before the experiment) and 2015–2016 (year of the experiment). All regressions are based on Equation (4). "Post" is a dummy equal to one if the evaluation is filled in the last 77% of the evaluation period. "year2015" is a dummy equal to one for the year of the experiment. "Female" is a dummy equal to one if the seminar teacher is a woman. All regressions include teacher fixed effects and control variables for students (student gender, age, whether the student is French, and variables to control for academic ability and admissions type). Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

logit estimations instead of OLS. Column (2) presents the results when using SET scores for both fall and spring semesters, and including student fixed effects instead of teacher fixed effects in the regressions. Column (3) presents the results when using SET scores for both fall and spring semesters, and adding student and teacher fixed effects. Columns (4) to (6) present the main specification (Column 7, Table 3) with standard errors clustered at the campus (column 4), teacher (column 5), or student level (column 6). Columns (7) to (9) present the results when using dummies equal to one if the SET overall satisfaction score is superior or equal to "average" (column (7)), "good" (column (8)) or "excellent" (column (9)). Results are similar to those presented in Table 3.

In the last column of Table D2, we present a placebo exercise where we run our main regression on scores in 2014–2015, one year before our experiment took place.

In Table D3, we present the results of difference-in-differences in control, treatment one and treatment two campuses with male teachers used as control. While the coefficients for post*Female are not significant (and negative) in control and treatment one groups, the coefficient is positive and significant in treatment two. These results are consistent with the ones presented in Table 3.

In Table D4, we test the robustness of our results when using the year before the experiment to build a control group. We run regressions of the following form:

Table D5
Effect of the treatment on different dimensions of teaching.

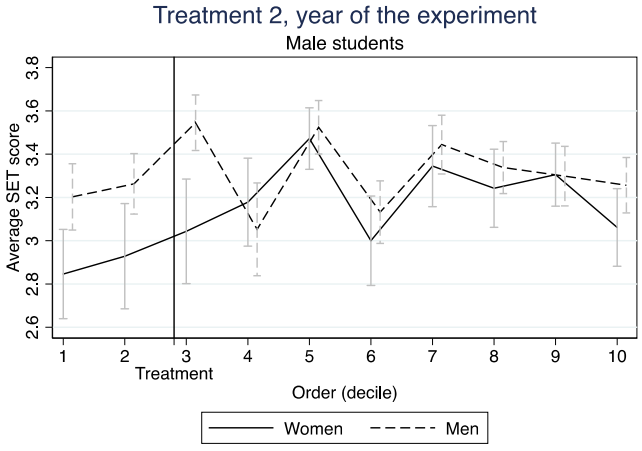
	(1) Preparation and organization	(2) Quality of instructional materials	(3) Clarity of course assessment criteria	(4) Usefulness of feedback	(5) Class leadership, animation skills	(6) Ability to encourage group work	(7) Availability and communication skills	(8) Ability to relate to current issues	(9) Contribution to intellectual development	(10) Effort invested by students	(11) Number of grades	(12) Deadline correction (oral exam)	(13) Deadline correction (written exam)
post	−0.075	0.13	0.0031	−0.014	−0.062	0.31**	0.021	0.078	−0.080	0.047	0.085	−0.041	0.0073
	(0.075)	(0.10)	(0.094)	(0.10)	(0.078)	(0.15)	(0.094)	(0.12)	(0.080)	(0.064)	(0.063)	(0.043)	(0.046)
post*T1	0.20**	−0.085	0.16	0.0063	0.19*	−0.11	0.052	0.0051	0.30***	0.0057	−0.043	0.10**	0.039
	(0.097)	(0.12)	(0.12)	(0.13)	(0.11)	(0.19)	(0.12)	(0.15)	(0.11)	(0.081)	(0.078)	(0.052)	(0.057)
post*T2	0.077	−0.094	0.058	0.064	0.11	−0.29*	0.022	−0.035	0.15*	0.0022	−0.075	0.037	−0.016
	(0.086)	(0.12)	(0.11)	(0.11)	(0.090)	(0.17)	(0.10)	(0.13)	(0.093)	(0.073)	(0.074)	(0.046)	(0.050)
post*female	0.079	−0.32**	−0.24*	−0.060	0.13	−0.057	−0.062	−0.055	0.044	0.048	−0.24**	0.028	−0.062
	(0.12)	(0.14)	(0.13)	(0.15)	(0.13)	(0.23)	(0.15)	(0.19)	(0.12)	(0.10)	(0.10)	(0.050)	(0.055)
post*female*T1	−0.15	0.32*	0.32*	0.24	−0.12	0.11	0.054	0.096	−0.088	−0.14	0.19	−0.025	−0.012
	(0.15)	(0.17)	(0.17)	(0.19)	(0.16)	(0.28)	(0.18)	(0.24)	(0.17)	(0.13)	(0.12)	(0.065)	(0.074)
post*female*T2	−0.018	0.41**	0.28*	0.19	−0.024	0.17	0.13	0.090	0.048	−0.053	0.27**	−0.019	0.064
	(0.14)	(0.17)	(0.16)	(0.18)	(0.15)	(0.25)	(0.17)	(0.21)	(0.15)	(0.12)	(0.12)	(0.057)	(0.062)
Observations	4472	4473	4472	4473	4472	4466	4470	4470	4473	4473	4471	4472	4463
Mean outcome	3.18	3.03	2.96	2.89	3.05	2.30	3.20	3.00	3.06	2.32	2.46	1.86	1.84
Sd outcome	0.87	1.02	0.98	1.04	0.96	1.36	0.92	1.14	0.96	0.63	0.83	0.35	0.37

Note: The dependent variable of each regression is specified in the column header. All regressions include teacher fixed effects and control variables for students (student gender, age, whether the student is French, and variables to control for academic ability and admissions type). Coefficients of T1 and T2 are absorbed by the teacher fixed effects. The questions are in Appendix A. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

$$SET_{s,te,t} = \beta_0 + \beta_1 * Woman_{te} + \beta_2 * post_t + \beta_3 * TreatYear_t + \beta_4 * Woman_{te} * post_t + \beta_5 * post_t * TreatYear_t + \beta_6 * post_t * TreatYear_t * Woman_{te} + \gamma * X_s + \delta * Z_{te} + \varepsilon_{s,te,t} \quad (4)$$

where $TreatYear_t$ is a dummy equal to one the year of the experiment, and $post_t$ is a dummy equal to one if the SET score was completed after the first 23.09% of the semester, i.e. after the email the year of the experiment or after the same point the year before (even though no email was sent then). We run by campus in columns (1) to (7), and by group in columns (8) to (10). We find that women's scores completed after the first 23.09% the year of the experiment are significantly higher in Paris and Le Havre (columns (6) and (7)), and, more generally, in treatment two (column (10)). These results are consistent with the ones presented in Table 3.

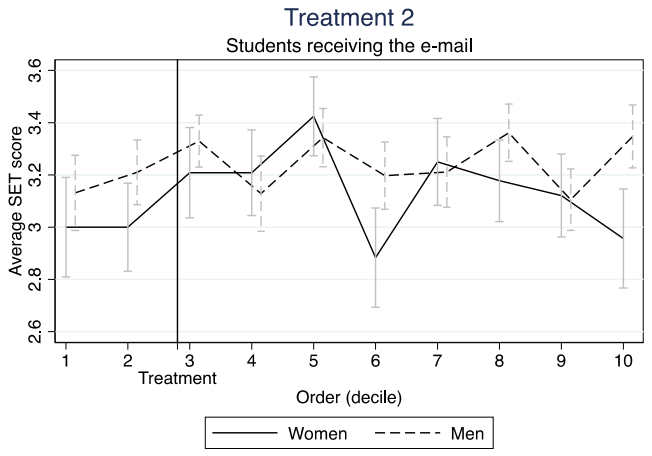
In Table D5, we present the effects of the treatments on the various dimensions of teaching. No clear pattern emerges.



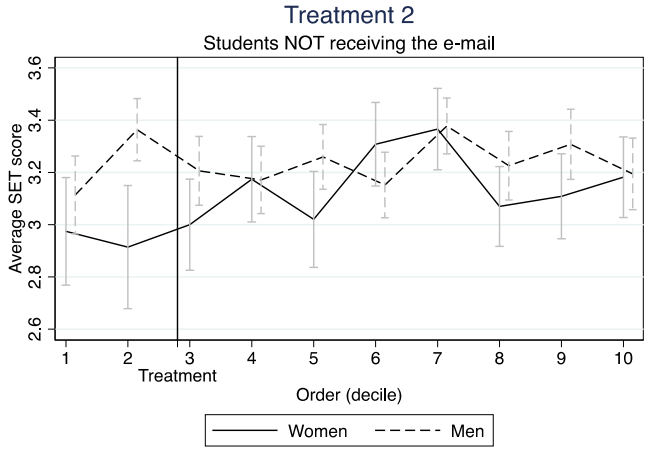
(a)

(b)

Fig. D1. Evolution of SET scores by teacher gender in treatment two campuses, male and female students separately. Note: each point indicates the mean SET score by gender (women-solid line or men-dashed line), and timing of evaluation (first decile completed, second decile completed, etc.). Subfigure (a) presents the evolution among male students in treatment two, while subfigure (b) presents the same evolution for female students. Segments indicate the confidence interval at 10%.



(a)



(b)

Fig. D2. Evolution of SET scores by teacher gender in treatment two campuses, students receiving the email or not. Note: each point indicates the mean SET score by gender (women-solid line or men-dashed line), and timing of evaluation (first decile completed, second decile completed, etc.). Subfigure (a) presents the evolution among students who received the email in treatment two, while subfigure (b) presents the same evolution for students who did not receive the email. Segments indicate the confidence interval at 10%.

References

Allcott, H., Rogers, T., 2014. The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. Am. Econ. Rev. 104 (10), 3003–3037.

Alesina, A., Carlana, M., Ferrara, E. L., Pinotti, P., 2018. Revealing Stereotypes: Evidence from immigrants in schools (No. w25333). National Bureau of Economic Research.

Asensio, J., Gómez-Lobo, A., Matas, A., 2014. How effective are policies to reduce gasoline consumption? Evaluating a set of measures in Spain. Energy Econ. 42, 34–42.

Arbuckle, J., Williams, B.D., 2003. Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. Sex Roles 49 (9–10), 507–516.

Bagues, M., Sylos-Labini, M., Zinovyeva, N., 2017. Does the gender composition of scientific committees matter?. Am. Econ. Rev. 107 (4), 1207–1238.

Bertrand, M., Duflo, E., 2017. Field experiments on discrimination. Handbook Econ. Field Exp. 1, 309–393.

Bertrand, M., Chugh, D., Mullainathan, S., 2005. Implicit discrimination. Am. Econ. Rev. 95 (2), 94–98.

Bohnet, I., Van Geen, A., Bazerman, M., 2015. When Performance trumps gender bias: joint vs separate evaluation. Manage. Sci. 62 (5), 1225–1234.

Boring, A., 2015. Gender biases in student evaluations of teachers. Document de travail OFCE 13.

Boring, A., 2017. Gender biases in student evaluations of teaching. J. Public Econ. 145, 27–41.

Boring, A., Ottoboni, K., Stark, P., 2016. Student evaluations of teaching (mostly) do not measure teaching effectiveness. ScienceOpen Research.

Collins, W.J., 2003. The labor market impact of state-level anti-discrimination laws, 1940–1960. ILR Rev. 56 (2), 244–272.

Collins, W.J., 2004. The housing market impact of state-level anti-discrimination laws, 1960–1970. J. Urban Econ. 55 (3), 534–564.

Dobbin, F., Kalev, A., 2016. Why diversity programs fail. Harvard Business Review (July-August), 52–60.

Dobbin, F., Kalev, A., 2018. Why doesn't diversity training work? The challenge for industry and academia. Anthropol. Now 10 (2), 48–55.

Fan, Y., Shepherd, L.J., Slavich, E., Waters, D., Stone, M., Abel, R., Johnston, E.L., 2019. Gender and cultural bias in student evaluations: why representation matters. PloS One 14, (2) e0209749.

Funk, P., Iriberri, N., Savio, G., 2019. When Margaret met Sally: Same-Sex Preferences in Academia when Female Instructors are Scarce.

Glover, D., Pallais, A., Pariente, W., 2017. Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores. Q. J. Econ. 132 (3), 1219–1260.

Goldin, C., Rouse, C., 2000. Orchestrating impartiality: the impact of "blind" auditions on female musicians. Am. Econ. Rev. 90 (4), 715–741.

Horne, Z., Powell, D., Hummel, J.E., Holyoak, K.J., 2015. Countering antivaccination attitudes. Proc. Natl. Acad. Sci. 112 (33), 10321–10324.

Ida, T., Ito, K., Tanaka, M., 2013. Using dynamic electricity pricing to address energy crises: evidence from randomized field experiments. 36th Annual NBER Summer Institute, Cambridge, MA, USA.

Kunze, A., Miller, A.R., 2017. Women helping women? Evidence from private sector data on workplace hierarchies. Rev. Econ. Stat. 99 (5), 769–775.

MacNell, L., Driscoll, A., Hunt, A.N., 2015. What's in a name: exposing gender bias in student ratings of teaching. Innov. Higher Educ. 40 (4), 291–303.

Mengel, F., 2020. Gender Bias in Opinion Aggregation. Available at SSRN 3572594.

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

Moss-Racusin, C.A., van der Toorn, J., Dovidio, J.F., Brescoll, V.L., Graham, M.J., Handelsman, J., 2014. Scientific diversity interventions. Science 343 (6171), 615–616.

Nyhan, B., Reifler, J., Richey, S., Freed, G.L., 2014. Effective messages in vaccine promotion: a randomized trial. Pediatrics 133 (4), e835–e842.

Nyhan, B., Reifler, J., 2015. Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information. Vaccine 33 (3), 459–464.

Oreopoulos, P., 2011. Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes. Am. Econ. J.: Econ. Policy 3 (4), 148–171.

Paluck, E.L., Green, D.P., 2009. Prejudice reduction: What works? A review and assessment of research and practice. Annu. Rev. Psychol. 60, 339–367.

Pope, D.G., Price, J., Wolfers, J., 2018. Awareness reduces racial bias. Manage. Sci.

Price, J., Wolfers, J., 2007. Racial Discrimination Among NBA Referees (No. w13206). National Bureau of Economic Research.

Rooth, D.O., 2010. Automatic associations and discrimination in hiring: real world evidence. Lab. Econ. 17 (3), 523–534.

Wagner, N., Rieger, M., Voorvelt, K., 2016. Gender, ethnicity and teaching evaluations: evidence from mixed teaching teams. Econ. Edu. Rev. 54, 79–94.

Yoeli, E., Hoffman, M., Rand, D.G., Nowak, M.A., 2013. Powering up with indirect reciprocity in a large-scale field experiment. Proc. Natl. Acad. Sci. 110 (Supplement 2), 10424–10429.