Reducing Discrimination in the Field:
Evidence from an Awareness Raising Intervention
Targeting Gender Biases in Student Evaluations of Teaching

Anne Boring
Erasmus School of Economics
& LIEPP (Sciences Po)
boring@ese.eur.nl

Arnaud Philippe
University of Bristol
arnaud.philippe@bristol.ac.uk

October 31st, 2019[1]

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.

Keywords: student evaluations of teaching, gender biases, field experiment

JEL: C93, I23, J71

[1] 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, the Milan Labour Lunch Seminar, and the 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, and National University of Singapore seminar participants. The authors would also like to thank the anonymous referees who provided very helpful suggestions. 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.

1. Introduction

Anti-bias awareness-raising campaigns are a common strategy that organizations use to reduce discrimination in the evaluations of employee performance and job candidate applications. These campaigns generally convey a normative message, often along the lines of "discrimination is wrong and unfair to employees and job candidates, as well as inefficient for the organization, so make sure that you do not discriminate". 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 may not believe that the normative message applies to their own behavior, because they may not realize that they are biased in their evaluation of others. Including information to make biases conscious may therefore be a necessary condition for anti-bias awareness-raising campaigns to have an effect.

This paper provides the result of a field experiment designed to test the impact of two types of awareness-raising campaigns: one with, and one without, information to generate bias awareness. We use the context of gender discrimination in student evaluations of teaching (SETs). The SET context resembles a common principle-agent problem. To make a personnel decision (such as a promotion), the employer (principle) needs information on an agent's performance. Individuals in the organization who directly observe the agent's actions (such as managers, coworkers or clients) are generally the ones who provide the information. If this information is biased, the principal's decision will also be biased. Therefore, eliminating biases in the agent's performance evaluations is important for the principal.

In the higher educational context, universities frequently rely on SET scores for retention and promotion decisions of instructors. For each course, students observe an instructor's actions, and provide information to university administrators about the instructor's performance. However, studies from different countries provide empirical evidence that students can be biased in their evaluations of female instructors: in the U.S. (Arbuckle & Williams, 2003; MacNell et al, 2014; Boring et al., 2016), France (Boring, 2017), the Netherlands (Wagner et al., 2016; Mengel et al., 2017), Switzerland (Funk et al., 2019), and Australia (Fan et al., 2019) for instance. This growing body of evidence on gender biases in SETs is driving some universities to reconsider their use.[2] Yet universities struggle

[2] For instance at the University of Southern California (https://academicsenate.usc.edu/teaching-evaluations-update/), the University of Oregon (https://provost.uoregon.edu/revising-uos-teaching-evaluations), and the Ontario Confederation of University Faculty Associations (https://ocufa.on.ca/blog-posts/briefing-note-report-of-the-ocufa-student-questionnaires-on-courses-and-teaching-working-group/). On September 9th 2019, the American Sociological Association also cited the body of research on biases in SETs in a formal statement to

to find alternatives to SETs. For instance, peer evaluations have drawbacks: they are time consuming and costly to implement and are also potentially biased. Eliminating evaluations altogether is not a viable alternative either: universities generally consider that having some information (even biased information) on an instructor's performance is better than not having any at all. Universities worldwide therefore have a clear interest in reducing gender biases in SET scores.

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" treatment—added information to trigger bias consciousness. It included the same statement as the purely normative treatment, plus information from a study on gender biases in SETs (Boring, 2015). The message informed students of the presence of gender biases in SET scores in previous years. It also contained precise information, 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-differences by teacher gender indicates 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

highlight some of the limitations of using SET scores as a measure of teaching effectiveness for promotion decisions in academia. This statement has been endorsed by 17 other scholarly associations. See www.asanet.org/studentevaluations.

scores of female teachers. This 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. The reduction of the gender gap following the informational email seems to be driven by male students increasing their scores for female teachers. 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 informative email. The effect of the informational treatment seems to have survived in the medium run, for the spring semester courses.

Finally, we find that the informational treatment had important spillover effects. On campuses treated by the informational email, we find an impact on both students who received the email and those who did not. Anecdotal evidence suggests that the email sparked conversations on campuses treated with the informational message, de facto treating other students. We find weak empirical evidence of a small delay in the effect on students who did not directly received the e-mail. 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 largely ignored the other treatment (i.e., the normative, blanket email sent by the administration with no precise information in it). The persistence of the effect in the medium run is more consistent with a learning explanation than with a purely behavioral one. However, since the email for the informational treatment 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).

This paper contributes to the literature on the efficiency of interventions designed to reduce discrimination. 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 & Green, 2009). There is still scant evidence that these strategies work in the field (Moss-Racusin et al. 2014; Bertrand & Duflo 2017), although some recent research suggests they might (Alesina et al., 2019). 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 & Kalev, 2016).

We extend the results from Pope et al. (2018), who study the impact of 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). The mechanism remains unclear. Indeed, the drop in discrimination may have been caused by changes in the behavior of the referees (those who discriminate) or the players (those who are being discriminated against).[3] If the drop was due to the referees, it may have been because: 1) they became aware of their own discriminatory behavior (learning effect through information); 2) they realized that they were under scrutiny (Hawthorne effect); or 3) the norm (views towards discrimination) became salient (normative effect). In our research, we focus exclusively on the change in behavior of the individuals who discriminate: our intervention only treats students, not teachers. 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 a drop-in discrimination.

Other research has focused on 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 & Miller, 2017; Bagues, et al., 2017), using joint evaluations (Bohnet et al., 2015), and anti-discrimination laws (Collins, 2003, 2004). In comparison to this literature, our paper shows that providing information on people's behavior–a strategy that is easier to implement than, for example, blind auditions–could be highly effective. In an example that is similar to our setting, committee members conducting interviews for European Research Council (ERC) grants have to watch a video[4] called "Recruitment Bias in Research Institutes", in which they learn about research results showing how such committee decisions can be gender biased.

While our approach has some limitations–in particular limited statistical power (due to the small number of treatment units) and difficulties to fully distinguish the mechanisms (learning vs priming or saliency)–it conveys important policy implications: purely normative

[3] While the referees may have stopped discriminating because the information made them aware of their own biases, the drop in discrimination may also have been caused by the players and coaches who adapted their game to avoid situations in which they would be discriminated against. For instance, Parsons et al. (2011) provide evidence that those who are discriminated against tend to change their behavior when they anticipate that they will be subject to discrimination. However, Pope et al. (2018) show that the behavioral changes are unlikely to be due to institutional changes made by the NBA following the media scrutiny, or by the firing of more (or hiring of less) biased referees.

[4] https://www.youtube.com/watch?v=g978T58gELo

awareness-raising campaigns may be ineffective, but 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.

Finally, this paper expands the literature on the effectiveness of various treatments aiming at inducing pro-social behavior through information. While a large share of this literature focuses on situations in which treated agents have some personal interest – benefiting from consuming less energy (Asensio et al., 2014; Ida et al., 2013; Allcott and Rogers, 2014), paying the cost of not consuming at a certain moment in time (Yoeli et al. 2013) – this paper focuses on a context where self-interest is limited.

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. for instance), in the fall semester of the 2015-16 academic year, on a cohort of 1,570 students. Several features of the university's first year undergraduate studies are useful for the experiment. First, all first-year students are required to 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 a week of work 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 located in seven separate campuses, with each campus focusing on a different geopolitical area. However, at the end of their three years of study, all students receive the same degree in social sciences. Paris is the largest campus. The other campuses are in Dijon, Le Havre, Menton, Nancy, Poitiers, and Reims.

Third, the administration makes it mandatory for students to complete their SETs online at the end of each semester. These 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.

Finally, in previous academic years (2008-2013), a study on SET scores in the Paris campus showed that students discriminated against female teachers (Boring et al., 2016; Boring, 2017), with male students being particularly biased in favor of male teachers. Overall satisfaction scores were biased, as well as scores on different dimensions of teaching.

2.2. Treatments

The administration formally agreed to let us run an experiment to test the two different treatments on the gender gap in SETs. We also received approval for our randomized controlled trial from J-Pal's Institutional Review Board (see appendix A).

Both treatments consisted in sending emails to students while they were completing their SETs. The normative treatment ("treatment one") encouraged students to avoid discrimination, especially gender discrimination (full text in appendix). 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 exact 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 might be biased too. The second email (full text in appendix) drew students' attention to the research

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[5], and that male students consistently gave higher overall satisfaction scores to male teachers. The email ended with the same normative statement as in the first treatment.

2.3. Design

In order to measure the effect 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 individual campus, they do not communicate between campuses, enabling us to send different emails to students on different campuses. Figure 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 1,509 evaluations for treatment one (95.8% response

[5] The final exam is designed by the main lecturer and is common to all students. This feature enables us to compare student learning across seminar groups. A different teacher from the one the student has in the seminar does the grading for the final exam. The grading is anonymous (double blind), and it takes place after students have completed their SETs for the semester. Seminar grades are not graded anonymously and are given to students before or during the period of time when students are completing their SETs. Seminar teachers are relatively free to design their own material. Seminar teachers design the exercises that will count for the seminar grades. Students' final grades for a course are a weighted average of the seminar grades and the final exam grade.

rate), 2,329 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 of 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, 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 (informative). The two emails were sent simultaneously on a Friday evening. In each treated campus, some evaluations were therefore completed before the treatment, and other evaluations after 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 first 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.

3. Identification strategies

The design of the experiment includes several 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:

$$SET_{s,te,t} = \beta_0 + \beta_1 * TT1 + \beta_2 * TT2 + \beta_3 * post_t$$
$$+\beta_4 * TT_1 * post_t + \beta_5 * TT_2 * post_t + \gamma * X_s + \delta_{te} + \varepsilon_{s,te,t} \qquad (1)$$

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 (student's 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.[6] 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 study on the same campuses

[6] 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. First, it drastically reduces the power of the regressions by introducing numerous fixed effects. Second, because students mainly fill all the SET of the semester on the same day, we could only measure the effect if we used both fall and spring semesters. The identification would come from the difference between the first and the second semester scores among students who filled their evaluations of the first 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.

as those who did. 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 do so by running regressions of the form:

$$SET_{s,te,t} = \beta_0 + \beta_1 * TT1 + \beta_2 * TC1 + \beta_3 * TT2 + \beta_4 * TC2 + \beta_5 * post_t + \beta_6 * TT_1 * post_t + \beta_7 * TT_2 * post_t + \beta_8 * TC_1 * post_t + \beta_9 * TC_2 * post_t + \gamma * X_s + \delta_{te} + \varepsilon_{s,te,t}$$

(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 emails one or 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:

$$SET_{s,te,t} = \beta_0 + \beta_1 * Woman_{te} + \beta_2 * post_t + \beta_3 * T1 + \beta_4 * T2 + \beta_5 * post_t * Woman_{te} + \beta_6 * post_t * T_1 + \beta_7 * post_t * T_2 + \beta_8 * Woman_{te} * T_1 + \beta_9 * Woman_{te} * T_2 + \beta_{10} * post_t * T_1 * Woman_{te} + \beta_{11} * post_t * T_2 * Woman_{te} + \gamma * X_s + \delta * Z_{te} + \varepsilon_{s,te,t}$$

(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 score). 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. Alternative specifications

The strategies presented above 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 which 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.[7] The results of these alternative strategies are in the appendix.

7 Another reason why those alternative models are not presented in the core of the paper is that they were not mentioned in the approval document we submitted to J-Pal's Institutional Review Board.

4. Main effects

4.1. Graphical evidence

In Figure 2, we present the evolution of the average overall satisfaction scores by teacher gender, groups (treatment two, treatment one, control), over time. We split the SET scores for each gender and each group into ten subgroups based on the time at which 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/largely treated group" and the last seven deciles constitute the post period. Subfigure 2a presents the evolution for male and female teacher in campuses included in treatment group two (2b and 2c are for treatment group one and the control group). In the last subfigure (2d), we present the same 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 SET scores. More importantly for the present study, Figure 2a indicates that female teachers' scores in treatment 2 campuses increased after the emails were sent. Male teachers' scores, however, do not seem to have been impacted by the treatment. In the treatment group one and the control group, SET scores of both female and male teachers do not seem to have changed after the e-mails. The evolution observed in treatment two the year of the experiment did not occur the year before (subgraph 2d).

The graphical evidence therefore suggests that treatment two increased women's SET scores. However, this pattern could be driven by the timing of the evaluation. Indeed, "good" and "bad" teachers could be evaluated at different points in time, and this evolution could drive the differences observed in Figure 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. Results are presented in Figure 3. The results are consistent with the ones presented in Figure 2: SET scores increased for female teachers 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, individual 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 of the regression presented in Equation (1) for women and men are shown in columns (1) and (2). The dataset is restricted to the students who received the emails (TT1 and TT2) and the students of the control group. 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. 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 among those who did not receive the emails but who studied 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 spillover effect of treatment two seems to be complete. Men's scores did not change significantly following treatment two. Once again, treatment one has had no impact on SET scores.

The fact that students who received and those who did not receive the email in treatment two campuses react similarly may be surprising. However, anecdotal evidence indicates that students extensively discussed the email in treatment two. 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.[8] The study was also published on the Facebook group of the campus feminist chapter, de facto treating other students on campus. These anecdotal pieces of evidence are consistent with the timing of male and female teachers' SET evolution in treatment 2 among students who received or did not received the email. While results are noisy because of small sample sizes, it seems that female teachers' SET scores given by

[8] 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."

students receiving the email may have 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 figure C1).

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 are in treated campuses). Results presented in columns (5) and (6) 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 further confirmed by triple-difference analyses. Column (7) shows the results of regressions including all overall satisfaction scores across all campuses. The results show that 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).

The effect presented in Table 3 is 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 e-mail. Whether the "informative treatment" would have had the same effect in the absence of such debates remains an open question.

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

In the appendix, we show that the results presented in Table 3 are robust to perturbations of the main specification (Table C1). Using ordered logit instead of ordinary least squares or controlling for student fixed effects or both teacher and student fixed effects (in a model including SET scores of both the fall and spring semesters) does not affect the results. 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 C1, columns 8-10). In comparison to men, women tend to get higher SET scores after the email were sent in treatment two (column 10) but not in control (column 8) nor treatment one (column 9) campuses. Second, using SET scores from 2014-2015 (the preceding year), we measure, for each campus (Table C2, columns 1-7) or group (Table C2, columns 8-10), the effects of completing evaluations in the last two weeks of the evaluation period[9], 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 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, despite the fact that 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.

In the appendix, we also show that we find no effect in a placebo exercise (Table C1, columns 7).

5. Mechanism

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. The email sent in treatment two explicitly referred to this difference among students. For this reason, two (non-exclusive) 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

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

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 if 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.

Second, we measure if treatment two impacted all female teachers or mainly benefited the better 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). 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 were the ones who 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.[10]

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. These results go against 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. The pre period becomes the reference for the evaluation filled both in the fall semester after the emails were sent (so for similar courses given by same teachers), and in the spring

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

semester (so with other courses and mostly other teachers). The identification of the effect in the spring semester is therefore weaker.

In appendix Table C3, we explore the effect of the treatments on the different teaching dimensions that the students also have to evaluate. Surprisingly, while only treatment two decreased the gender gap on overall satisfaction scores, the two different treatments seem to have the same effect on the teaching dimensions, and may have reinforced gender stereotypes[11]. Women's scores in "quality of instructional materials" or "clarity of course assessment" are significantly better after both treatments, while all teachers' scores in "contribution to intellectual development" are significantly better after the treatment. Other teaching dimensions do not seem to be impacted.

6. Conclusion

What constitutes an effective way to educate students about their own biases is still very much of an open research question. Nonetheless, several researchers from universities that use SET scores in promotion decisions have reached out to us to know how these results could apply to their contexts. Given the specificities of our field experiment context, the policy advice we can give is 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.

How should universities engage students about discrimination in SET scores? Some instructors have told us that they worry that if they are the ones who try to encourage students to treat all professors equally, the intervention may backfire against them individually. For instance, one female instructor wrote to us the following: "I first fear that it would encourage students to think that I have a problem and score me lower because I can't handle being female in a male world. Then I fear that if I did tell them about the bias that I would do it "wrong" leading to encouraging or prompting them to act on it rather than try to avoid it." To

[11] Boring (2015) finds that the dimensions that students value in men and women tend to correspond to gender stereotypes. For example, women get better scores in teaching dimensions such as course preparation and organization, while men get better scores in "contribution to intellectual development" and class leadership skills.

avoid such uncomfortable and potentially counterproductive situations for instructors, it may be necessary for the intervention to be carried-out by the administration, and not the instructors being evaluated. Furthermore, the administration must beware to avoid a potential counterproductive activation of stereotypes through the anti-bias intervention (Dobbin & Kalev, 2018).

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. Indeed, a poorly designed message can be ineffective. The results may partly 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.[12]

[12] Horne et al. (2015) study information campaigns designed to reduce anti-vaccination beliefs,and find that campaigns that attempt to refute vaccination myths are inefficient, sometimes even counter-productive—generating more people to hold anti-vaccination beliefs (Nyhan et al., 2014; Nyhan & Reifler, 2015). However, Horne et al. (2015) find evidence that campaigns providing factual evidence on the negative consequences of communicable diseases (such as measles) on children can efficiently lead parents to vaccinate their children.

References

Allcott, H., & Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, *104*(10), 3003-37.

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 Economics*, *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?. *The American Economic Review*, *107*(4), 1207-1238.

Bertrand, M., & Duflo, E. (2017). Field experiments on discrimination. *Handbook of Economic Field Experiments*, *1*, 309-393.

Bertrand, M., Chugh, D., & Mullainathan, S. (2005). Implicit discrimination. *The American Economic Review*, *95*(2), 94-98.

Bohnet, I., Van Geen, A., & Bazerman, M. (2015). When Performance Trumps Gender Bias: Joint vs. Separate Evaluation. *Management Science*, *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. *Journal of Public Economics*, *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 Review*, *56*(2), 244-272.

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

Cavalcanti, T., & Tavares, J. (2016). The Output Cost of Gender Discrimination: A Model-based Macroeconomics Estimate. *The Economic Journal*, *126*(590), 109-134.

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. *Anthropology 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. *The Quarterly Journal of Economics*, *132*(3), 1219-1260.

Goldin, C., & Rouse, C. (2000). Orchestrating Impartiality: The Impact of" Blind" Auditions on Female Musicians. *The American Economic Review*, *90*(4), 715-741.

Horne, Z., Powell, D., Hummel, J. E., & Holyoak, K. J. (2015). Countering antivaccination attitudes. *Proceedings of the National Academy of Sciences*, *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. *Review of Economics and Statistics*, *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. *Innovative Higher Education*, *40*(4), 291-303.

Mengel, F., Sauermann, J., & Zölitz, U. (2017). Gender bias in teaching evaluations. *Journal of the European Economic Association* (forthcoming).

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. *American Economic Journal: Economic Policy*, *3*(4), 148-171.

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual Review of Psychology*, *60*, 339-367.

Parsons, C. A., Sulaeman, J., Yates, M. C., & Hamermesh, D. S. (2011). Strike three: Discrimination, incentives, and evaluation. *The American Economic Review*, *101*(4), 1410-1435.

Pope, D. G., Price, J., & Wolfers, J. (2018). Awareness reduces racial bias. *Management Science*.

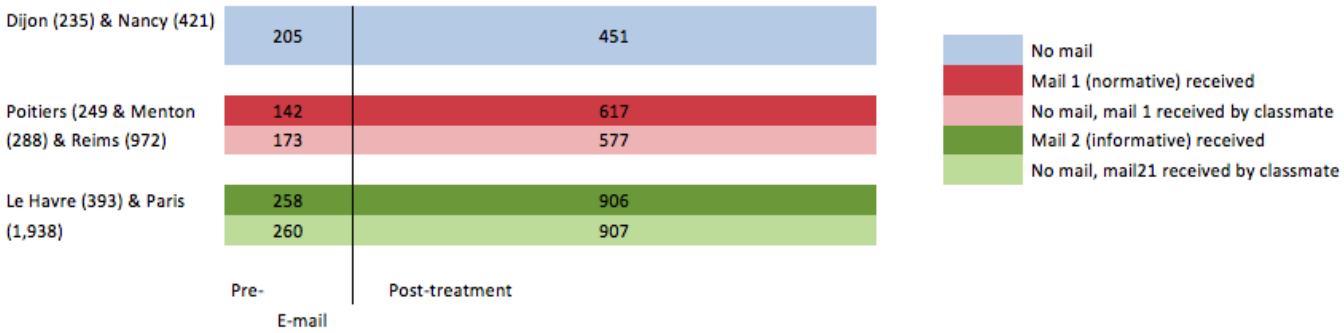
Price, J., & Wolfers, J. (2007). *Racial Discrimination Among NBA Referees* (No. w13206). National Bureau of Economic Research.

Rooth, D. O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, *17*(3), 523-534.

Wagner, N., Rieger, M., & Voorvelt, K. (2016). Gender, ethnicity and teaching evaluations: Evidence from mixed teaching teams. *Economics of Education Review*, *54*, 79-94.

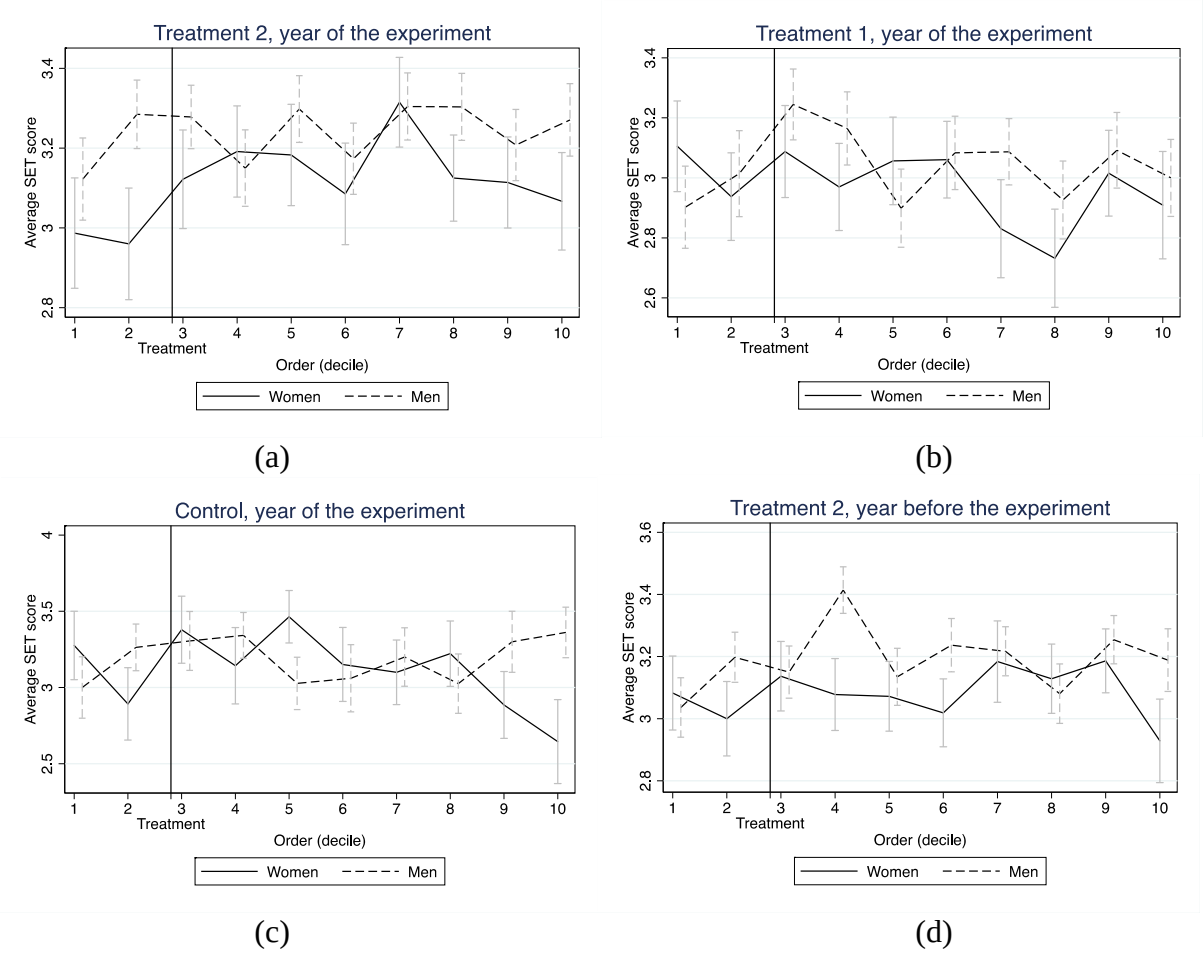
Yoeli, E., Hoffman, M., Rand, D. G., & Nowak, M. A. (2013). Powering up with indirect reciprocity in a large-scale field experiment. *Proceedings of the National Academy of Sciences*, *110*(Supplement 2), 10424-10429.

Figure 1: Design of the experiment



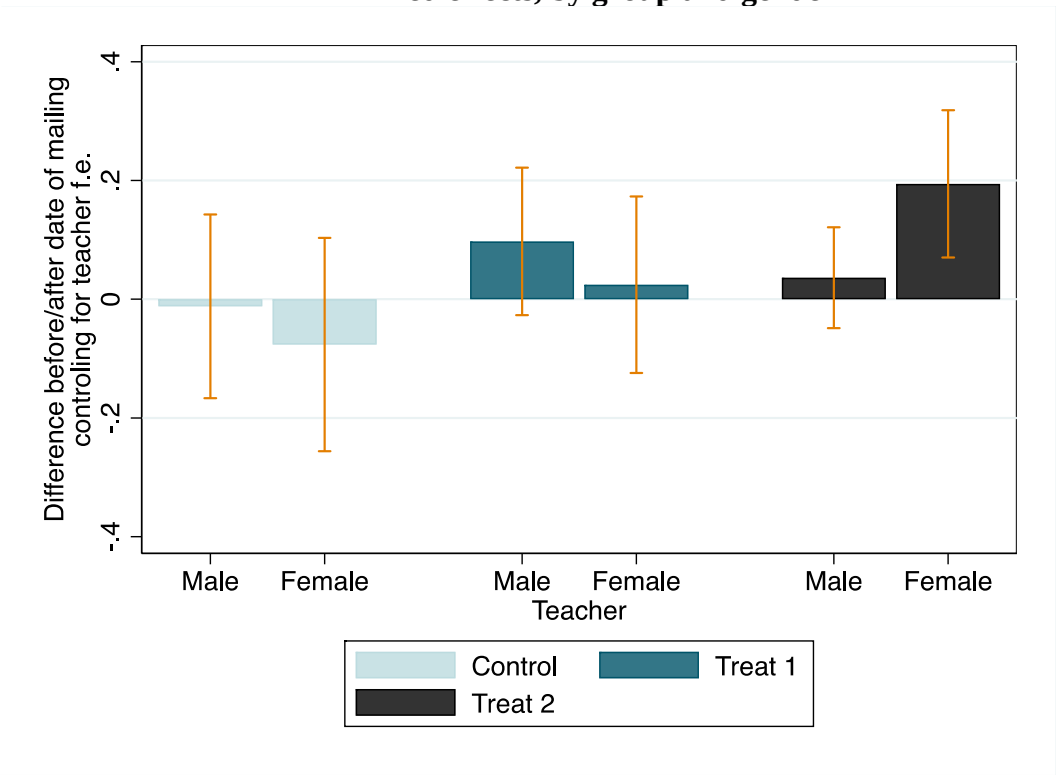
Note: Each color represents a different group. The numbers indicate the number of observations in the dataset (i.e.. the number of evaluations). The black bar indicates the moment when the administration sent the e-mails. The campuses included in each group are indicated on the left hand side.

Figure 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 filled, second decile filled… Subgraphs (a) (b) and (c) present the evolution in treatment two, treatment one 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%.

Figure 3: Difference in SET score before and after mail after controlling for teacher fixed effects, by group and gender



Note: each bar indicates the difference in SET scores before and after e-mails for different groups (control, treatment one, or treatment two) and a different gender (women or men). Segments indicate the confidence interval at 5%.

Table 1. Descriptive statistics on students and teachers

	Mean	S.d.
Students		
Share of women	.60	.49
Age	18.17	.79
Continuous assessment (seminar) grade	139.86	22.46
Final exam grade	116.81	34.35
Share of students with French citizenship	.73	.44
Share of students admitted through specific procedure	.10	.31
Share of students admitted through entry exam (French high school)	.46	.50
Share of students admitted through international procedure	.32	.47
Share of students admitted through dual degree with a foreign university	.08	.27
Share of students admitted through dual degree with a French university	.02	.16
Share of students enrolled in a regular degree	.79	.41
Share of students enrolled in a dual degree with a foreign university	.10	.30
Share of students enrolled in a dual degree with a French university	.11	.31
Teachers		
Share of women	.39	.49
Share of "excellent" overall satisfaction scores	.40	.49
Share of "good" overall satisfaction scores	.38	.49
Share of "average" overall satisfaction scores	.15	.36
Share of "insufficient" overall satisfaction scores	.06	.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

Table 2. Balancing checks

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Student female	Final exam grade	Continuous assessment grade	French citizenship	Age	Entry exam waived	Entry exam	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	4,496	4,473	4,496	4,496	4,496	4,496	4,496	4,496

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. They are not significant.

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	1,025	1,542	1,727	2,746	1,727	2,746	4,473
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			
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, variables to control for academic ability, and variables to control for 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.

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	1,766	2,707	2,154	2,319	2,369	2,104	8,655

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 variable T1 and T2 are absorbed by the teacher fixed effects in all columns.

Appendix A. The Two Emails Sent

Mail 1 :

Cher(e) étudiant(e),

Les évaluations en ligne des enseignements sont ouvertes depuis le lundi 23 novembre 2015. Le remplissage de ces évaluations fait partie de vos obligations de scolarité. Comme il vous l'a été précisé dans l'email signalant l'ouverture des évaluations en ligne, les informations que vous complétez sont lues par les enseignant-es et utilisées avec beaucoup d'attention par la Direction des études et de la scolarité afin de préparer chaque rentrée universitaire. Vos appréciations permettent en particulier à la direction de Sciences Po d'améliorer, en lien étroit avec les équipes pédagogiques, la qualité de nos formations.

Il convient à ce titre de rappeler que les évaluations ne doivent porter que sur la qualité des enseignements et qu'elles ne doivent pas être influencées par des facteurs tels que le sexe, l'âge ou l'origine ethnique des enseignant(e)s. Nous vous demandons de faire tout particulièrement attention à ces questions de discriminations afin d'éviter que, par exemple, les enseignantes soient systématiquement moins bien notées que leurs homologues masculins en raison de biais ou de stéréotypes de genre.

Nous vous prions de croire, cher(e) étudiant(e), à l'assurance de nos sentiments les meilleurs.

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,

Hélène Kloeckner
Chargée de la communication interne / Référente égalité femmes-hommes

SciencesPo
Direction de la communication / Secrétariat général
27 rue Saint-Guillaume 75337 Paris cedex 07 France
T. +33 (0)1 45 49 59 86 / **M.** +33 (0)6 73 76 32 96
helene.kloeckner@sciencespo.fr
www.sciencespo.fr

Mail 2 :
Cher(e) étudiant(e),

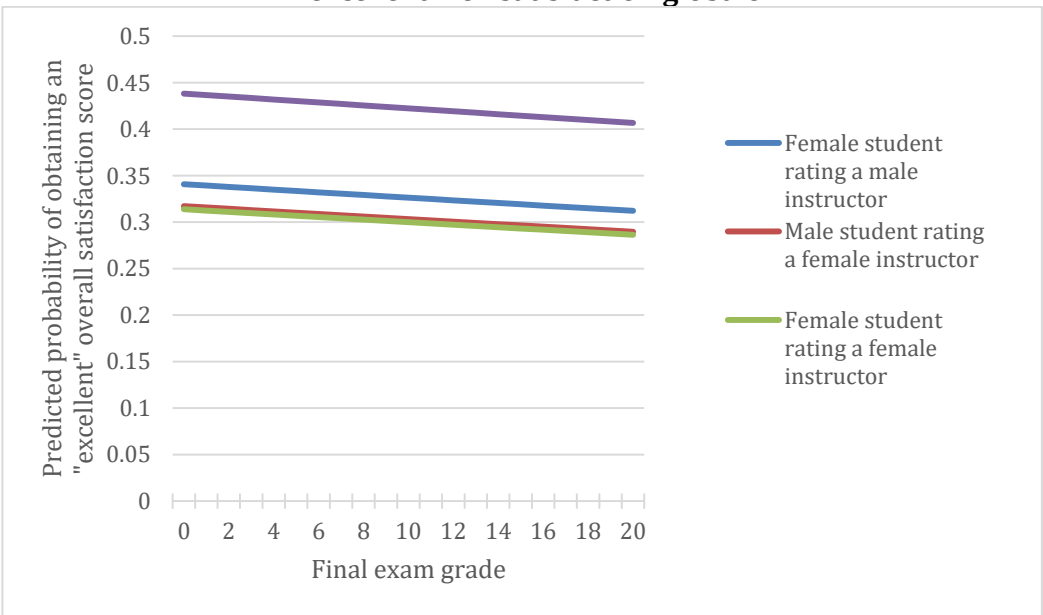
En cette période d'évaluation des enseignements nous souhaitons attirer votre attention sur les résultats d'une recherche récente menée à Sciences Po mettant en évidence un biais discriminatoire à l'encontre des femmes enseignant les conférences de méthode pour les modules fondamentaux de première année.

Il s'avère en effet qu'à résultat égal aux examens, les élèves tendent à moins bien noter les enseignantes. Cet écart s'observe en particulier de la part des élèves hommes bien que les élèves femmes présentent également un biais. Ces écarts ne semblent pas justifiés par d'autres mesures de la qualité d'un enseignement, telle que la capacité d'un(e) enseignant(e) à faire réussir ses élèves aux examens de fin de semestre.

Prenons par exemple le cas d'élèves obtenant 13,5 de moyenne en conférence de méthode et 12 à l'examen final (ce qui correspond aux moyennes observées sur la période d'étude 2008-2013, tous modules fondamentaux confondus). Pour ces élèves, les enseignantes ont 30% de chances d'obtenir un score de « satisfaction globale » qualifié d'excellent, quel que soit le sexe de l'étudiant (et à caractéristique d'enseignement constant, par exemple le jour et l'heure du cours). En revanche, pour ces mêmes notes en contrôle continu et à l'examen final, les enseignants obtiennent un score de satisfaction globale qualifié d'excellent dans 33% des cas s'ils sont évalués par une femme et même dans 42% des cas s'ils sont évalués par un homme. Cela signifie qu'à résultats des élèves égaux, les enseignantes obtiennent d'excellentes évaluations environ 19% moins souvent que leurs homologues masculins (compte tenu de la proportion moyenne d'élèves femmes et hommes). Ces différences sont statistiquement significatives.

Par ailleurs, quelle que soit la note obtenue à l'examen final, les élèves hommes évaluent systématiquement mieux les enseignants hommes, comme le montre le graphique ci-dessus.

Graphique : Corrélation entre note à l'examen final et probabilité prédite d'un score « excellent » en satisfaction globale



Enfin, les résultats de cette étude suggèrent que les élèves appliquent des stéréotypes de genre dans la façon dont ils répondent aux questions plus précises (notamment la question portant sur la qualité de l'animation et celle portant sur la contribution au développement intellectuel).

Au regard de ces résultats, il convient de rappeler que les évaluations ne doivent porter que sur la qualité des enseignements et qu'elles ne doivent pas être influencées par des facteurs tels que le sexe, l'âge ou l'origine ethnique des enseignant(e)s. Nous vous demandons de faire tout particulièrement attention à ces questions de discriminations afin d'éviter que, par exemple, les enseignantes soient systématiquement moins bien notées que leurs homologues masculins en raison de biais ou de stéréotypes de genre.

Nous vous prions de croire, cher(e) étudiant(e), à l'assurance de nos sentiments les meilleurs.

Dear Student,

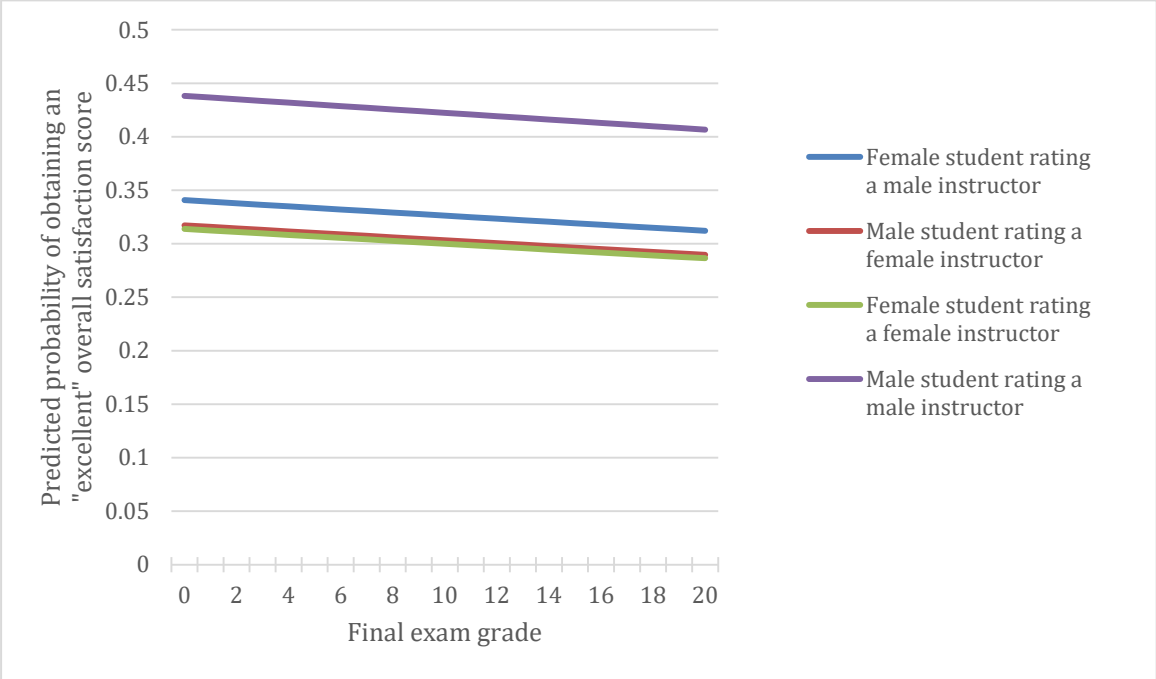
In this period of student evaluations of teaching (SET), we would like to bring your attention to the results of a recent study which suggests the existence of gender biases against female instructors of first year undergraduate seminars (i.e. the *conférences de méthode*) for all fundamental courses.

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,

Hélène Kloeckner
Chargée de la communication interne / Référente égalité femmes-hommes

SciencesPo
Direction de la communication / Secrétariat général
27 rue Saint-Guillaume 75337 Paris cedex 07 France
T. +33 (0)1 45 49 59 86 / **M.** +33 (0)6 73 76 32 96
helene.kloeckner@sciencespo.fr
www.sciencespo.fr

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. Additional Material

This appendix presents additional results mentioned in the paper.

Tables C1 and C2 present some robustness checks of our main results. In Table C1 we present perturbations of our main specification. Column (1) presents the results when using order 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 results when using dummies equal to one if the SET overall satisfaction score is superior or equal to "average" (column (4)), "good" (column (5)) or "excellent" (column (6)). Columns (7) to (9) present the results of difference-in-differences in control, treatment one and treatment two campuses when male teachers are used as control. Results are similar to those presented in Table 3.

In the last column of Table C1, 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 C2, 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:

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

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 C3, we present the effects of the treatments on the various dimensions of teaching. No clear pattern emerges.

Table C1. Robustness checks

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Ordered logit	Fall and spring semester	Fall and spring semester	Dummy			Difference in difference: Male as control			Placebo: year before the experiment
		Student fixed effects	Teacher and student fixed effects	1 vs 2/3/4	1/2 vs 3/4	1/2/3 vs 4				
	All	All	All	All	All	All	Control	Treatment 1	Treatment 2	All
post	0.012	-0.048	-0.067	0.011	0.0087	0.0058	0.022	0.11*	0.048	-0.016
	(0.23)	(0.12)	(0.090)	(0.021)	(0.036)	(0.051)	(0.074)	(0.061)	(0.043)	(0.086)
post*T1	0.34	-0.034	0.14	0.047	0.042	0.016				0.10
	(0.28)	(0.15)	(0.11)	(0.029)	(0.046)	(0.063)				(0.11)
post*T2	0.094	0.029	0.020	0.012	0.014	0.0061				0.17*
	(0.26)	(0.14)	(0.10)	(0.025)	(0.041)	(0.057)				(0.094)
post*female	-0.26	-0.17	-0.065	-0.029	-0.048	-0.030	-0.078	-0.079	0.18**	0.011
	(0.37)	(0.14)	(0.11)	(0.039)	(0.053)	(0.074)	(0.12)	(0.094)	(0.077)	(0.13)
post*female*T1	0.027	-0.058	0.0035	-0.010	0.0053	0.030				-0.11
	(0.45)	(0.19)	(0.15)	(0.050)	(0.071)	(0.092)				(0.16)
post*female*T2	0.82*	0.36**	0.29**	0.040	0.14**	0.10				-0.21
	(0.43)	(0.17)	(0.14)	(0.046)	(0.065)	(0.087)				(0.15)
Observations	4,473	8,655	8,630	4,473	4,473	4,473	654	1,503	2,316	4,398

Note: all regressions include teacher fixed effects and control variables for students (student gender, age, whether the student is French, variables to control for academic ability, and variables to control for admissions type). Coefficients of variables T1 and T2 are absorbed by the teacher fixed. Columns 1, 4, 5, and 6 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 7 contains all the evaluations completed in the fall semester 2014-2015.Columns 8,9,10 present the diff in diff (Male teachers as control groups) for control, treatment one and treatment two campuses.

Table C2. 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	1,543	720	4,268	1,289	2,594	4,988

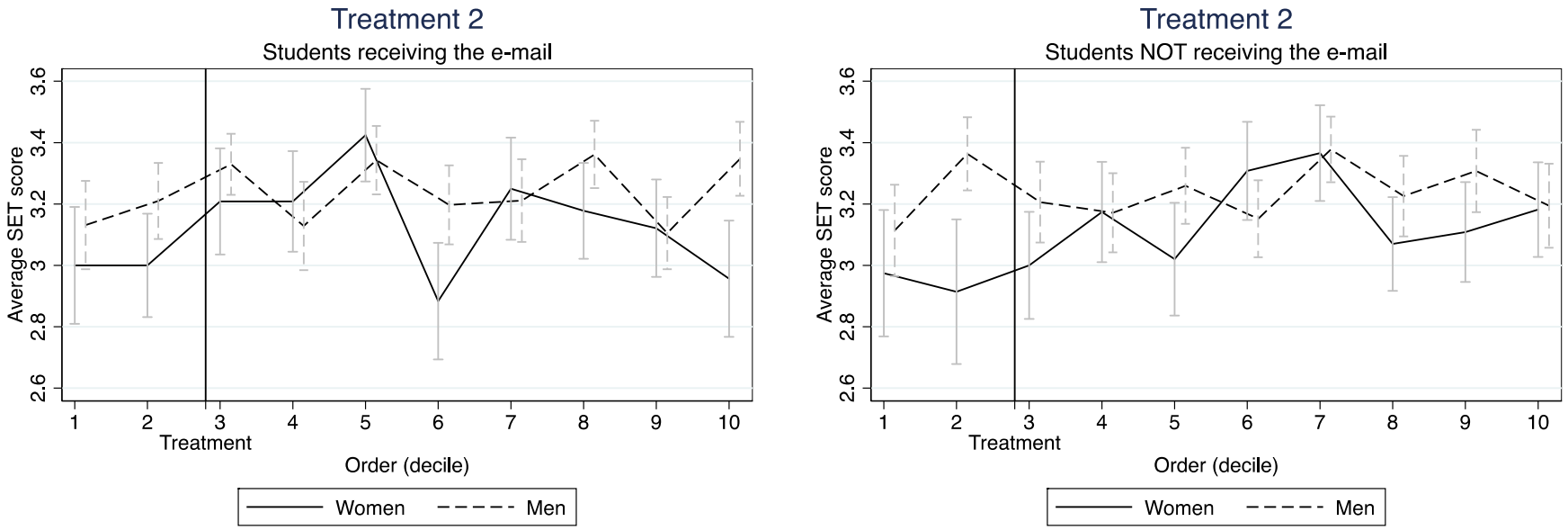
Note: The sample is composed of evaluation filled in the first semester of school year 2014-2015 (before the intervention) and 2015-2016 (year of the experiment). All regressions are based on Equation 4. "Post" is a dummy equal to one if the evaluation is filled in the last 77% of the year. "year2015" is a dummy equal to one 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, variables to control for academic ability, and variables to control for admissions type).

Table C3. Effect of the treatment on different dimensions of teaching

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	Preparation and organization	Quality of instructional materials	Clarity of course assessment criteria	Usefulness of feedback	Quality of animation	Ability to encourage group work	Availability and communi-cation skills	Ability to relate to current issues	Contribution to intellectual development	Investment	Number of grade	Deadline correction (oral exam)	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	4,472	4,473	4,472	4,473	4,472	4,466	4,470	4,470	4,473	4,473	4,471	4,472	4,463

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, variables to control for academic ability, and variables to control for admissions type). Coefficients of T1 and T2 are absorbed by the teacher fixed effects. In order to simplify the table, coefficients of variables TT1 and TT2 are not presented in columns 3 and 4. They are not significant.

Figure C1: Evolution of SET scores by teacher gender in treatment 2 campuses



(a) (b)

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 filled, etc.). Subfigure (a) presents the evolution among students who received the email in treatment two, while figure (b) presents the same evolution for students who did not received the email. Segments indicate the confident interval at 10%.