Reducing Discrimination through Norms or Information: Evidence from a Field Experiment on Student Evaluations of Teaching

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

Arnaud Philippe
Institute for Advanced Study in Toulouse - Toulouse School of Economics
arnaud.philippe@iast.fr

November 7th, 2017[1]

ABSTRACT

We conduct a field experiment to assess the impact of two different interventions designed to reduce gender biases in student evaluations of teaching (SET). In the first intervention, students received a normative statement by email, essentially reminding them that they should not discriminate in SETs. In the second intervention, the normative statement was augmented with precise information on how other students in the exact same situation had discriminated against female teachers in the past. While the pure normative statement had no significant impact on SETs, the informative statement appears to have reduced gender biases against female teachers. This effect 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 of AFSEE 2016 and Advances with Field Experiments 2017 conference participants, as well as CREST, IAST, Erasmus University Rotterdam and George Mason University seminar participants. 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

Public policies have made a priority of reducing widespread discrimination[2] over the past decades. A popular strategy has been to try eliminate biases by changing individuals' beliefs, tastes or values, through awareness-raising campaigns. These campaigns generally carry the normative message that people should not discriminate, because discrimination is wrong or unfair. Do such strategies work? Maybe not. Biases (that lead to discriminatory behaviors) tend to be unconscious (Bertrand et al., 2005; Rooth, 2010; Oreopoulos, 2011). Discriminating individuals may therefore not feel that such normative messages apply to their own behavior, because they lack information on their own biases. Promoting a normative narrative about how discrimination is wrong can even be counter-productive, for instance if it includes a "blaming and shaming" approach (Galinsky & Moskowitz, 2000; Dobbin & Kalev, 2013). On the other hand, a more informative intervention designed to make individual biases conscious may work.

In this paper, we present the results of a field experiment in which we tested the impact of a normative and an informative message to reduce discrimination. Our research focuses on reducing gender biases in student evaluations of teaching (SETs). Multiple studies conducted in different contexts have shown that students tend to discriminate against female teachers (MacNell et al, 2014; Boring et al., 2016; Wagner et al., 2016; Boring, 2017; Mengel et al., 2017).[3] The experiment consisted in nudging students through two emails designed to eliminate biases in evaluations.

A first email encouraged a group of students to be careful not to discriminate in SETs (the "normative" treatment). We designed this strategy to resemble ubiquitous awareness-raising campaigns. Another group of students received an email that included the same message, plus information from a working paper (Boring, 2015) on gender biases in SETs (the "informative" treatment). This second message informed students that, in previous years,

[2] Just to cite a few papers or books that study discrimination in different contexts: wages (Becker, 1957; Arrow, 1973; Hamermesh & Biddle, 1993; Blau & Kahn, 2017; Weichselbaumer & Winter-Ebmer, 2005), hiring (Black, 1995; Phelps, 1972; Goldin & Rouse, 2000; Bertrand & Mullainathan, 2004; Riach & Rich 2006; Booth & Leigh 2010), job promotions (Lazear & Rosen, 1990); housing (Ondrich et al., 1999; Ewens, Tomlin & Wang, 2014); the sharing economy (Ge et al., 2016; Edelman et al., 2017); health (Balsa & McGuire, 2001); sports (Parsons et al., 2011); crime and prison sentences (Eberhardt et al., 2004; Philippe, 2016); car sales (Ayres & Siegelman, 1995); mortgage lending (Ladd, 1998), etc.

[3] Universities often rely on SET scores for promotion and tenure decisions of teachers. They must therefore avoid making unfair and inefficient personnel decisions, by ensuring that students do not discriminate. When biases are present, universities that persist in using SETs for personnel decisions have two main options. They could try to de-bias scores *ex post*. However, research suggests that the extent of gender discrimination is highly context-dependent, making it almost impossible for universities to do so in practice (Boring et al., 2016). Another solution is to eliminate biases *ex ante*, so that the scores objectively reflect the actual quality of instruction. In this paper we test this *ex ante* strategy.

students of the same university, in the same context, expressed gender biases in SETs. The design of the informative treatment made the students explicitly aware of the presence of these biases.

To test the impact of these messages, we take advantage of the existence of seven different campuses in the university to create a difference-in-differences setting. The students of two campuses were considered as controls. They did not receive any email during the three-week mandatory online evaluation period. Three other campuses were treated with the normative message, and the two remaining campuses were treated with the informative message. The emails were sent after some students had already completed their SETs. 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 for the students who completed their SETs after the emails were sent.

The results of the analysis show that the second treatment was effective in the sense that it prompted students to increase their evaluations of female teachers, thus reducing the gender gap in SET scores. However, we find little evidence that the purely normative statement had any significant impact on SET scores. The spillover effect within campus is extremely high, especially in the information treatment campuses. One of the reasons why the informative treatment may have worked was because it sparked discussions on gender discrimination among students, as anecdotal evidence suggests.

These results show that the efficiency of de-biasing interventions is highly dependent on a message's content. Public policies aimed at reducing discrimination must therefore be carefully designed.

The paper is organized as follows. Section 2 describes the institutional setting. Section 3 presents the experiment, and section 4 the identification strategy. The main results are in section 5. Section 6 discusses the possible mechanisms. Section 7 concludes.

2. Institutional setting

At this French university specialized in social sciences, all first year students are required to follow several mandatory courses in history, macroeconomics, microeconomics, political institutions (law), political science, and sociology. The experiment took place in the fall semester of the 2015-16 academic year. We then study the long run impact on the spring semester courses. 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 administration requires students to

complete their SETs online each semester. These SETs remain anonymous to the teachers, who cannot trace back SET scores to individual students. Students who share the same main lecture all take the same final exam. The final exam grade counts for one third of the final grade. The main lecturers design the final exam. The seminar grade counts for the other two thirds of the final grade. The seminar teachers design the exercises that will count in the seminar grades. Final exams are graded anonymously (double blind) after students are completing 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.

Undergraduate students are located in seven separate campuses. At the end of their three years of study, they all receive the same degree in social sciences. Paris is the largest campus. The other campuses are in Dijon (Central and Eastern European campus), Le Havre (Europe-Asia campus), Menton (Middle-Eastern and Mediterranean campus), Nancy (German and European campus), Poitiers (Spanish, Portuguese and Latin-American campus), and Reims (African and North American campuses). The main difference in the campuses has to do with the languages taught in each campus, and the fact that international students are admitted to any of the campuses outside Paris (depending on the language they wish to study). While students know each other quite well within each individual campus (especially the smaller campuses), they generally do not communicate between campuses.

In this university, male teachers receive higher overall satisfaction scores than female teachers. In past academic years (2008-2013), students from the Paris campus have been shown to discriminate against female teachers (Boring et al., 2016; Boring, 2017), with male students being particularly biased in favor of male teachers. Overall satisfaction scores are biased, as well as scores on different dimensions of teaching (preparation and organization of courses, class leadership skills, etc.).

3. Presentation of the experiment

The university's administration has been seeking ways to reduce these biases to avoid penalizing female teachers. Within the context of a European-funded (FP7) project called *Effective Gender Equality in Research and the Academia* (EGERA), the university accepted our research proposal, to test scientifically what type of intervention would be more efficient. The administration formally agreed to let us run an experiment to test the two different treatments.

We also received approval for our randomized controlled trials from J-Pal's Institutional Review Board (see appendix). Our research protocol explicitly stated the hypothesis we wanted to test: because students may be unaware of their biases in SETs, signaling to students that these biases exist could help to reduce them.

3.1. Treatments

The goal of the experiment was to test the effect of two different treatments on the gender gap in SETs. Both treatments consisted in sending emails to students while they were completing their SETs.

The first 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. The email 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."*

The statement did not include any trigger to suggest that discrimination might have occurred in this precise context. This treatment resembles many anti-discrimination campaigns, whose main message is that "individuals should not discriminate". If biased individuals are not conscious that they discriminate, this type of message is likely to be ineffective.

The second treatment ("treatment two") added precise information to the normative statement, by explicitly telling students that students had been gender-biased in the past, in the exact same context. By making students identify with other (biased) students in the same context, this treatment may reveal to students that they too might be biased. 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" 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 are unrelated to student performance on the final exam, and that male students consistently give higher overall satisfaction scores to male teachers. The email ended with the same normative reminder as in email 1.

3.2. Design

In order to measure the effect of these two treatments, we take advantage of the fact that the university has separate campuses. The design of the experiment is presented in Figure 1. The first treatment group includes students from three campuses: Menton with 102 students, Poitiers with 86 students and Reims with 337 students. We assigned the normative email (treatment 1) to this group of students. The second treatment group includes students from two other campuses: Le Havre with 131 students and Paris with 657 students. We assigned the informative email (treatment 2) to students from this second group. The other two campuses, Dijon with 101 students and Nancy with 155 students, are the control group campuses: students did not receive any emails.

The response rate is high, given that completing SETs is mandatory. The database includes a total of 1,509 evaluations for the treatment one campuses (95.8% response rate), 2,329 evaluations for the treatment two campuses (98.5% response rate), and 656 evaluations for the control group campuses (85.4% response rate).

We sent the emails to only half of the students of each campus that were part of treatments one and two. The students who received emails were randomly selected before the beginning of the evaluation process. The different groups and the sample sizes are

summarized in Table 1. We use the following notations: group C is the control group (all students in Dijon and Nancy); group TT1 (treatment treated 1) includes all students who received the *normative* email; TC1 (treatment control 1) includes all students who did not receive the email, but who were on the campuses that were treated with the normative email; TT2 and TC2 are similar to TT1 and TC1 but for the campuses where students received the *informative* email.

We sent the two emails after some students had started completing their SETs, within the three-week time span when students were required to rate their teachers. The emails were sent when roughly one fifth of the evaluations had been completed: 20.9% in treatment one and 22.2% in treatment two. The two emails were sent simultaneously. Some evaluations were therefore completed before the treatment, and some other after the treatment in each treated campus. The university's gender equality officer sent the emails.

3.3. Data

We ran our experiment in the fall semester of the 2015-16 academic year, on a cohort of 1,570 students. Table 2 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, because admission to the first year at this university can only occur right after high school. Students tend to receive higher continuous assessment grades (nearly 14 out of 20 on average), and lower final exam grades (11.7 out of 20 on average). Many students are French: 73% have French citizenship (including some who are dual citizens of another country). Finally, 32% of students were admitted through the international procedure (they went to high school abroad), 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 are dual degree students.

Among teachers, 39% are women. Most teachers obtain overall satisfaction scores that students qualify as "excellent" (39%) or "good" (40%). Very few overall satisfaction scores are "insufficient" (only 6%), while a slightly larger share of students give "average" overall satisfaction scores (15%). Overall satisfaction scores tend to be higher in history (3.21 on average), than in political institutions (3.09 on average) and microeconomics (3.08 on average).

4. Identification strategies

The main objective of the experiment is to evaluate the effects of the two treatments on SET scores. The design of the experiment includes three features that enable us to use difference-in-differences analyses. First, some campuses are treated while others are control. Second, on the treated campuses, the emails were sent after some students had already completed their evaluations, generating a pretreatment period. Third, only half of the students (random draw) received emails on the treatment campuses. If students who received an email on the treatment campuses discuss the emails with students who were not intended to be treated, then there might be spillover effects. The identification strategy enables us to measure these spillover effects.

4.1. Measuring the effect of the treatment

We first eliminate any spillover effects by using an analysis that only includes 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 more or less affected by spillover effects, depending on the magnitude of these effects.

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

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

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; and Z_{te} are controls for teacher characteristics.

We test whether the emails are effective by observing whether they cause a reduction in the gender gap in SET scores. Our variables of interest are β_4, which measures the effect of the normative treatment, and β_5, which measures the effect of the informative treatment.

The controls and fixed effects determine the variations and the sample used to identify the effect. A first (simple) option is to include a limited set of control variables (student's

gender, age, grade). In theory, the difference-in-differences structure is enough to identify the effect and, in this case, we can use all the observations from semester one and two. However, the measurement could be biased if the pre-treatment SETs concern teachers who share some characteristics or are completed by students who also share some characteristics.

A second option is to include teacher fixed effects in the first specification. The advantage of using this specification is that it overcomes the potential bias due to the heterogeneity of SET timing based on teachers' characteristics. An advantage of using teacher fixed effects is that male and female teachers do not need to be of the same quality. Our preferred specifications are therefore regressions including teacher fixed effects.

Including teacher fixed effects overcomes the potential bias due to correlations between timing and teachers' characteristics, but not the potential bias due to correlations between timing and students' characteristics. It could be interesting to include student fixed effects in the main specification. However, including students fixed effects presents several limitations. First, it drastically reduces the power of the regressions by introducing numerous fixed effects. Second, 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 and the identification will come from the difference between the first and the second semester among students who filled their evaluation of the first semester before the treatments. For this reason, using student fixed effects is only valid if the gender gap would have been similar in the campuses during the two semesters in the absence of the treatment. This hypothesis is stronger than the one needed when we only use the fall semester – i.e. "the evolution of the gender gap is similar before and after the emails in the different campuses" – and could not be tested. Another drawback of this strategy is that it mainly measures the medium run effect of the treatment. For those reasons we will only use regressions with student fixed effects as robustness checks.

4.2. Spillover effect of the treatment and net effect

In the strategy presented in the previous section, we focused on the effect of the treatment on the students who received the email in comparison to the students of the control group. However, we are able to measure spillover effects through the two groups of students (TC1 and TC2) who did not receive an email, but who study on the same campuses 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. We do so by running regressions of the form:

$$SET_{s,te,t} = \beta_0 + \beta_1 * TT1 + \beta_2 * TC1 + \beta_3 * TT2 + \beta_4 * TC2 + \beta_5 * post_t + \beta_6 * TT_1 * post_t + \beta_7 * TT_2 * post_t + \beta_8 * TC_1 * post_t + \beta_9 * TC_2 * post_t + \gamma * X_s + \delta * Z_{te} + \varepsilon_{s,te,t} \quad (2)$$

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 have 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 is total. We run equation (2) separately for female and male teachers. Following the discussion presented in section 4.1., the specifications included in the core of the article will include teacher fixed effects.

Lastly, we can 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 have a very important spillover effect, and if TT1/TC1 and TT2/TC2 are very close.

4.3. Triple difference in difference

Another possibility is to mix the two difference-in-differences strategies into one single triple differences strategy. As the results are harder to read when using a triple difference in difference, and in order to limit the number of coefficients, 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 1 and 2 (respectively).

4.4. Balancing checks

Our main identification assumption relies on the fact that the differences between students who complete SETs or teachers who are evaluated before and after the emails were sent are similar across groups. Fixed effects – teacher, group or student – help to partly relax this hypothesis. However, it is also possible to partially test for this assumption directly by running balancing checks on observable characteristics of the evaluations. We run regressions of the form:

$$Charact_{s,te,t} = \beta_0 + \beta_1 * T1 + \beta_2 * T2 + \beta_3 * post_t$$
$$+\beta_4 * T_1 * post_t + \beta_5 * T_2 * post_t + \gamma * X_s + \delta * Z_{te} + \varepsilon_{s,te,t} \qquad (3)$$

where $Charact_{s,te,t}$ is the observable characteristics of the evaluation filled at time *t* by student *s* on teacher *te* and the rest of the variables are similar to equation 1 and 2. Regressions based on equation (3) are run on male and female teachers separately. As we are interested into the correlation between characteristics and timing, we do not need to differentiate between students who received the emails and students around them (i.e. we do not need to differentiate between TC1 and TT1 or TC2 and TT2).

Results are presented in Table 2. All the regressions include teacher fixed effects. We mainly run balancing checks on the characteristics of the students. However, grades could be viewed as depending on both students' and teachers' characteristics.

5. Main effects

5.1. Graphical evidence

Before presenting the formal regression model and our main results, we present graphical evidence that captures the idea of the main treatment effects. We are interested in the evolution of the difference before and after the treatment. In Figure 2, we present the average overall satisfactions scores by sex of the teachers and groups, comparing before and after the emails were sent. On average, men's SET scores are greater than women's SET scores in the three groups. Figure 2 shows more specifically that female teachers' scores increase after the emails were sent ("Female teacher post") on the treatment two campuses. In both the control campuses and the treatment one campuses, female teachers' average overall satisfaction scores drop after the emails, suggesting that treatment one may have had no impact on female teachers' scores on average. Male teachers' average overall satisfaction scores remain relatively constant in the control group (there is a slight drop). Their scores tend appear to increase, however, in the treatment one campuses following the email. There is also a smaller increase in male teachers' scores after the emails in the treatment two campuses.

The graphical evidence therefore suggests that treatment two increases women's SET scores. Furthermore, the emails may have had an impact on men's scores in both treatment campuses.

5.2. Main results

We first measure the effect of the two treatments on the overall satisfaction scores using difference-in-differences analyses, as presented in sections 4.1 and 4.2 (for the spillover effects). Table 4 presents the main results. Regressions include controls for students' observable characteristics (age, whether the student is French, student's continuous assessment and final exam grades, students' average grades in other courses, and admission type), as well as teacher fixed effects to control, among other things, for teachers' teaching styles.

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

(students from the Dijon and Nancy campuses, who received no email). The results show that treatment two increases female teachers' SET scores (column 1). After the mailing campaign, the informative treatment induces a significant increase of 0.26 point for women. Treatment one has no significant effect and the effects of treatment one and two are not statistically different (see the p-value of the test of equality between the effect of treatments one and two). 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 increases 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 study on the treatment two campuses. The difference between the effects on these two different 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 do not change significantly following treatment two, and treatment one once again seems to have no impact on SET scores of women and men.

We hypothesized that the informative message was efficient in our context because it prompted students to become conscious of the biases that they might apply when completing SETs. We have anecdotal evidence that the informative emails generated discussions among students, at least on one of the treatment two campuses. At the end of the year (after students were done completing evaluations for the spring semester courses), we sent an email to students on the Le Havre campus (where one of us teaches a main lecture), asking whether they had discussed the content of the email with one another (were there any discussions at all)? Or did the email remain largely unnoticed/unread? Some students mentioned that they did indeed discuss the email with other fellow students. For instance, one female student said: "I remember this email very well because it created a long debate/discussion among my group of friends and I." The feminist chapter of the campus seems to have especially taken-up the issue, including on the Facebook page of students on campus. These discussions among students explain why we observe a complete spillover effect within campus.

Given this evidence of within campus spillover effects, we measure the effect of the treatments without distinguishing between students directly treated (those who received the email) or indirectly treated (those who did not receive the email but who are in treated campuses) in columns (5) and (6). Treatment two has 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 a statistically significant impact on either women or men.

These results are further confirmed by triple-differences 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 receive higher overall satisfaction scores after emails are sent (the coefficient on $post_t * T_2 * Woman_{te}$ shows a statistically significant increase of 0.28 point).

5.3. Robustness checks

To further test the robustness of our findings we measure the effect of the treatments using different models or different variables, in Table 5. Columns (1) to (3) present the results when we use a dummy equal to one if SET scores are good or excellent. Columns (4) to (6) present the main results while using ordered logit regressions instead of OLS. In columns (7) to (12) the sample size is extended to the SET filled during both the fall and spring semester. Columns (7) to (9) present the results when regressions include controls for student fixed effects instead of teacher fixed effects. Lastly, in columns (10) to (12), regressions include both student and teacher fixed effects. For each robustness check, we first present the results of the difference-in-differences on women and men in two separate columns and the last column presents the results of the triple difference in difference. All the results presented in table 5 give the net effect of the treatments: TC1 and TT1 as well as TC2 and TT2 are not distinguished. Results are similar to the ones observed in Table 4. In the four robustness checks, we find that treatment one has no effect while treatment two increases SET scores for women. Moreover, the magnitudes of the results are similar to those observed in table 4.

In another series of robustness checks, we include only one of the treated campuses at a time with the two control campuses (Table 6). The results are not consistent across the treatment one campuses. Indeed, in Poitiers, the overall satisfaction scores for both women and men increase following the emails (weakly significant for men). The results are consistent for the treatment two campuses however: the scores of women (and only women) significantly increase in both the Paris and Le Havre campuses following the email. The effect appears to be larger in Le Havre. This is due to the fact that women in the pre email period on this campus have much lower SET scores compared to men, and compared to teachers in the Paris campus.

6. Mechanism

6.1. Effect of the treatment by student and teacher gender

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, which generated the higher overall satisfaction scores for male teachers. The email sent in treatment two explicitly referred to this difference among students. We therefore check whether male students, who were more specifically targeted in the email of treatment two, react more.

We start by presenting graphs of the overall SET score by sex of the teachers and groups – as in figure 2 – for male and female student separately. The graphs are presented in figure 3.a. (male students) and 3.b. (female students). Among female students, the satisfaction scores for men slightly increase in all groups. Women's scores decrease in groups C and T1, and slightly increase in T2. Among male students, two evolutions are striking: male teachers have significantly better scores after the treatment 1 while female teachers have significantly better scores after treatment 2. The graphical evidence presented in figures 3 suggest that the effect of the treatment mainly comes from male students. They also suggest that the normative treatment may have increased the gender discrimination among male students.

In order to further investigate these hypotheses, we run the difference-in-differences as well as the triple differences on male and female students separately. Results are presented in Table 7. All regressions include teacher fixed effects as well as controls for students' characteristics. The results confirm that the effect of the treatment comes from male students. Both the results from difference-in-differences and triple differences show that female teachers receive higher overall satisfaction scores after treatment two because male students increase their evaluations for women. The results also suggest that male students increase the scores they give to female teachers following treatment one. However, this result is only observed while using difference-in-differences (column 1), and it is not confirmed by triple differences (column 5).

In contradiction with figure 3a, the regressions do not support the idea that male students tend to give better evaluations to male teachers after treatment one. Lastly, the emails

appear to have had no statistically significant impact on female students' evaluations of both female and male teachers.

6.2. Other dimensions of the evaluations

Until now, we only focused on overall satisfaction scores. However, the complete evaluations are composed of thirteen other questions. These questions are more precise and cover students' opinions on teachers' effectiveness on different dimensions of teaching. 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.

We explore the net effect (with the full sample and T1 and T2 instead of TT1 and TT2) of the treatment on these dimensions using triple differences following equation (3). Results are presented in Table 8. All regressions include teacher fixed effects.

Surprisingly, while only treatment two decreased the gender gap on overall satisfaction scores, the two different treatments seem to have the same effect, and may have reinforced gender stereotypes. 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.

6.3. Heterogeneity of the effects

We document the heterogeneity of the effects of the treatments along three dimensions: teachers' quality, students' quality and fields. Results are presented in Table 9 using triple difference in differences analyses.

First, columns (1) and (2) include results of regressions separating the better teachers from the other 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 in the campus). Although results are weakly significant, the higher quality female teachers are the ones who are especially benefitting from the higher overall satisfaction scores with treatment two.

Second, we measure whether "good" students react differently (columns (3) and (4)). We define "good" students as those who get final grades above the median on campus. We use regressions similar to the ones used for "good teachers". This analysis does not yield statistically significant results, suggesting that both types of students may be increasing the overall satisfaction scores of female teachers.

Lastly, we separately measure the effect of the treatments by course – economics, history and law (fall semester courses). The number of female teachers varies largely. While they are 51% of the teachers in economics, they are only 36% in history and 29% in law. Sample sizes are small and all the results are non significant. However, the coefficient for law is slightly smaller than the ones for economics and history.

1. Conclusion

One of the main conclusions that can be drawn from this field experiment is that the content of an awareness-raising campaign is important. Indeed, a poorly designed message can be ineffective or, worse, actually generate an increase in discrimination. The persistence of discrimination may be surprising given the millions of dollars spent every year by firms and governmental agencies on diversity training, as well as by governmental 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.[4] Recent research has also studied strategies to counter "alternative facts", finding that trying that presenting voters with the true facts can actually backfire and generate extra political support for the politicians who promoted alternative facts (Barrera et al., 2017).

We focus on a strategy to this research provides evidence from the field that "de-biasing" strategies might be effective. Indeed, while a few laboratory experiments using implicit association tests have studied such strategies (Paluck & Green, 2009), there has been scant evidence that these strategies work in the field (Moss-Racusin et al. 2014; Bertrand & Duflo 2017). Our experiment contributes to the research asking whether it is possible to reduce discrimination by changing individuals' beliefs, tastes or values through awareness-

[4] 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.

raising campaigns. Other research has focused mainly on reducing discrimination by studying changes in the settings or rules in which discriminatory decisions are made. For instance, Goldin and Rouse (2000) show how having "blind" auditions helped to reduce discrimination against female candidates in orchestras. Blind performance evaluations are hard to apply in most workplace contexts however. Other researchers have studied the impact of increasing the number of women in hiring committees or in the hierarchy (Kunze & Miller, 2014) on discrimination in women's employment opportunities or promotions. Another solution is to implement gender quotas on hiring and promotion committees. This solution has yielded mixed results: under some circumstances, adding women to an evaluating committee can actually be more harmful to the women being evaluated (Bagues, et al., 2017). In yet another solution, Bohnet et al. (2015) suggest that conducting joint evaluations may reduce biases compared to separate evaluations (for which evaluators may rely more on group stereotypes). However, other research suggests that evaluators may shift their decisions of the criteria that they find to be the most important, in order to fit their evaluations with their gender biases, a phenomenon called casuistry (Norton et al., 2004). Furthermore, providing evaluators with information on a candidate's past performance does not completely eliminate discrimination in stereotypically male fields (Reuben et al., 2014). Finally, others have studied the impact of anti-discrimination laws (Collins, 2003, 2004).

Discrimination persists in contexts in which economic agents (employers, customers, students, etc.) evaluate the quality of women's and minorities' work or qualifications. Personnel decisions based on discriminatory behavior are therefore economically inefficient. Employers may make suboptimal hiring decisions based on stereotypes, for instance hiring less qualified individuals because their identity corresponds to the stereotype of the dominant group in a field (for instance, favoring less qualified men in science over more qualified women, Reuben et al., 2014). Individuals who expect to be discriminated against may also choose to underinvest in their education or development of skills, because they believe that discrimination will cause the returns to their human capital investment to be low. Reducing discrimination could therefore boost the economy (Cavalcanti & Tavares, 2016). For instance, the think tank *France Stratégie* estimates that reducing workplace discrimination against women and minorities could lead to a €150 billion increase in France's GDP over a twenty-year period, i.e. approximately a 7% increase in GDP (Bon-Maury et al., 2016). Furthermore, the growth of the share economy in recent years has increased the relevance of this field of research. Several recent studies have found evidence of large discrimination in this area. For instance, Edelman et al. (2017) find evidence of racial discrimination against African-

Americans making housing requests on Airbnb. Ge et al. (2016) also find evidence of discrimination against African-American users of Lyft and Uber. As online rating platforms develop, reducing discrimination in evaluations is becoming increasingly important.

References

Arrow, K. (1973). The theory of discrimination. *Discrimination in labor markets*, *3*(10), 3-33.

Ayres, I., & Siegelman, P. (1995). Race and gender discrimination in bargaining for a new car. *The American Economic Review*, 304-321.

Bagues, M., Sylos-Labini, M., & Zinovyeva, N. (2017). Does the Gender Composition of Scientific Committees Matter?. *The American Economic Review*, *107*(4), 1207-1238.

Balsa, A. I., & McGuire, T. G. (2001). Statistical discrimination in health care. *Journal of Health Economics*, *20*(6), 881-907.

Barrera, O., Guriev, S., Henry, E., & Zhuravskaya, E. (2017). *Facts, Alternative Facts, and Fact Checking in Times of Post-Truth Politics* (No. 12220). CEPR Discussion Papers.

Becker, G. S. (1957). The Economics of Discrimination. *University of Chicago Press Economics Books*.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review*, *94*(4), 991-1013.

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.

Black, D. A. (1995). Discrimination in an equilibrium search model. *Journal of Labor Economics*, *13*(2), 309-334.

Blau, F. D., & Kahn, L. M. (2017). The Gender-Wage Gap: Extent, Trends, and Explanations. *Journal of Economic Literature*, *(55)*3, 789-865.

Bohnet, I., Van Geen, A., & Bazerman, M. (2015). When Performance Trumps Gender Bias: Joint vs. Separate Evaluation. *Management Science*, *62*(5), 1225-1234.

Booth, A., & Leigh, A. (2010). Do employers discriminate by gender? A field experiment in female-dominated occupations. *Economics Letters*, *107*(2), 236-238.

Bon-Maury, G., Dherbecourt, C., Flamand, J., Gilles, C., Bruneau, C., Diallo, A., & Trannoy, A. (2016). *Rapport : Le coût économique des discriminations*, France Stratégie, Documentation Française.

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*.

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., & Roberson, Q. M. (2013). The Origins and Effects of Corporate Diversity Programs. *Oxford Handbook of Diversity and Work*.

Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing black: race, crime, and visual processing. *Journal of Personality and Social Psychology*, *87*(6), 876.

Edelman, B., Luca, M., & Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, *9*(2), 1-22.

Ewens, M., Tomlin, B., & Wang, L. C. (2014). Statistical discrimination or prejudice? A large sample field experiment. *Review of Economics and Statistics*, *96*(1), 119-134.

Edelman, B., Luca, M., & Svirsky, D. Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. *American Economic Journal: Applied Economics*.

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology*, *78*(4), 708.

Ge, Y., Knittel, C. R., MacKenzie, D., & Zoepf, S. (2016). *Racial and gender discrimination in transportation network companies* (No. w22776). National Bureau of Economic Research.

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

Hamermesh, D. S., & Biddle, J. E. (1994). Beauty and the Labor Market. *The American Economic Review*, *84*(5), 1174-1194.

Ladd, H. F. (1998). Evidence on discrimination in mortgage lending. *The Journal of Economic Perspectives*, *12*(2), 41-62.

Lazear, E. P., & Rosen, S. (1990). Male-female wage differentials in job ladders. *Journal of Labor Economics*, *8*(1, Part 2), S106-S123.

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.

Norton, M. I., Vandello, J. A., & Darley, J. M. (2004). Casuistry and social category bias. *Journal of Personality and Social Psychology*, *87*(6), 817.

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.

Ondrich, J., Stricker, A., & Yinger, J. (1999). Do landlords discriminate? The incidence and causes of racial discrimination in rental housing markets. *Journal of Housing Economics*, *8*(3), 185-204.

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.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The American Economic Review*, *62*(4), 659-661.

Philippe, A. (2017). *Gender disparities in criminal justice* (No. 17-762). Toulouse School of Economics (TSE).

Reuben, E., Sapienza, P., & Zingales, L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, *111*(12), 4403-4408.

Riach, P. A., & Rich, J. (2006). An experimental investigation of sexual discrimination in hiring in the English labor market. *Advances in Economic Analysis & Policy*, *5*(2).

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.

Weichselbaumer, D., & Winter-Ebmer, R. (2005). A meta-analysis of the international gender wage gap. *Journal of Economic Surveys*, *19*(3), 479-511.

Figure 1: design of the experiment

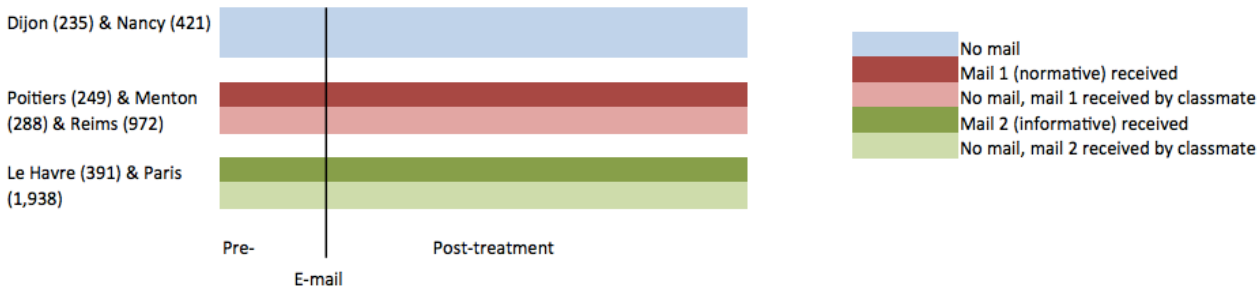


Figure 2: Mean SET scores by period, teacher gender, and groups

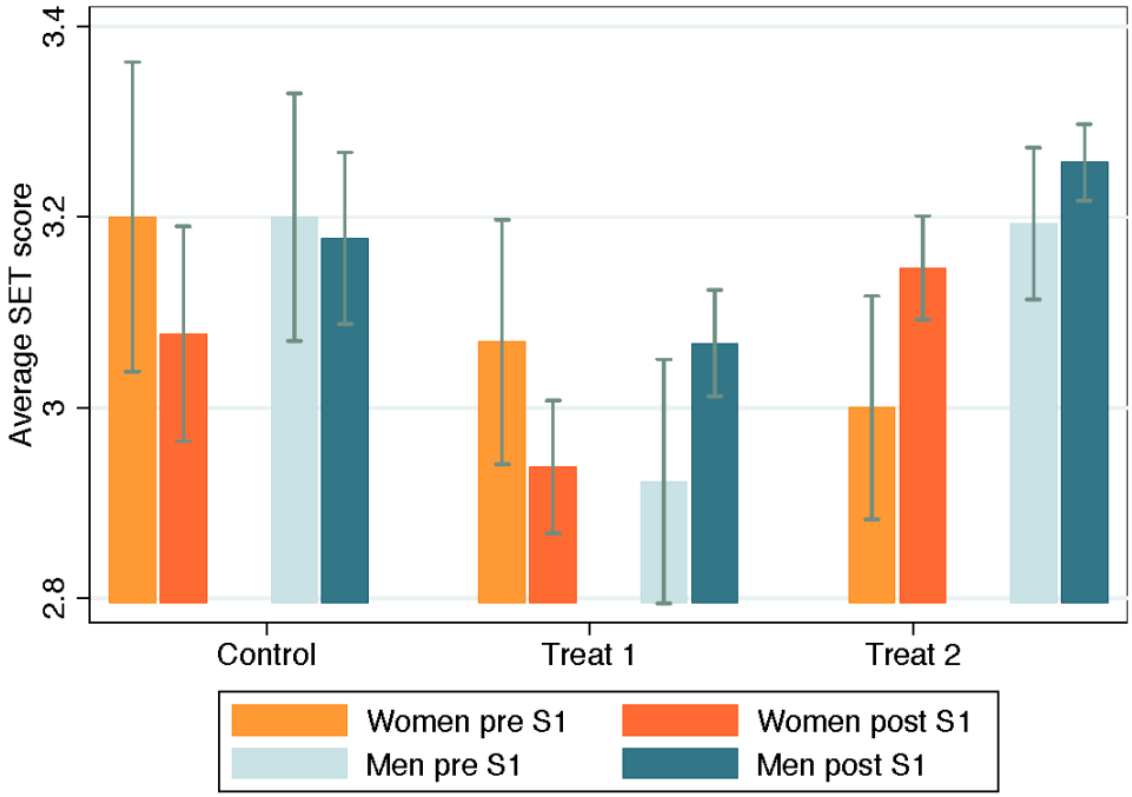


Figure 3.a.: Mean overall satisfaction scores by male students, before vs. after, by teacher gender and groups

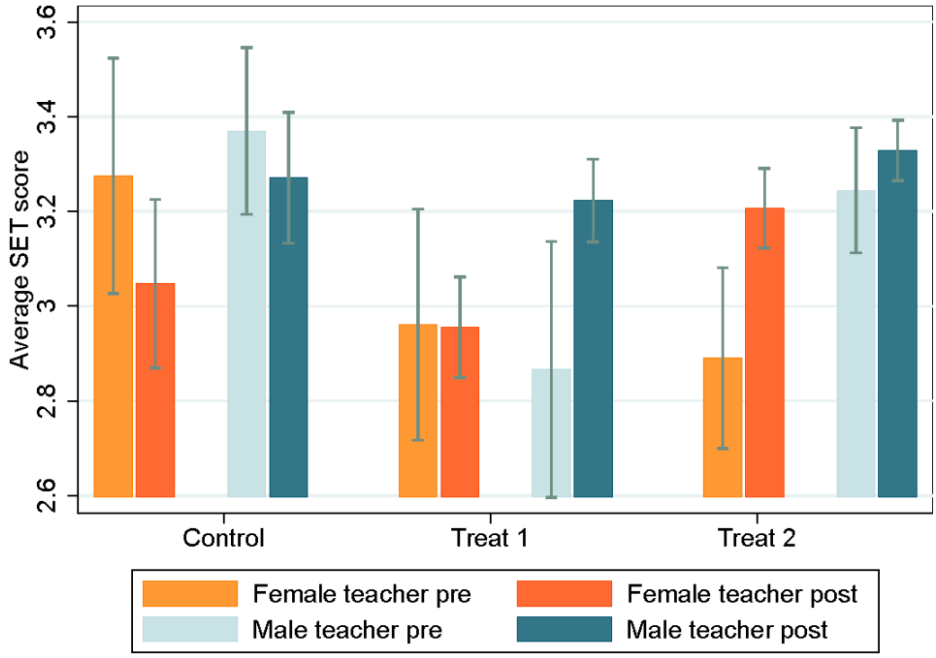


Figure 3.b.: Mean overall satisfaction scores by female students, before vs. after, by teacher gender and groups

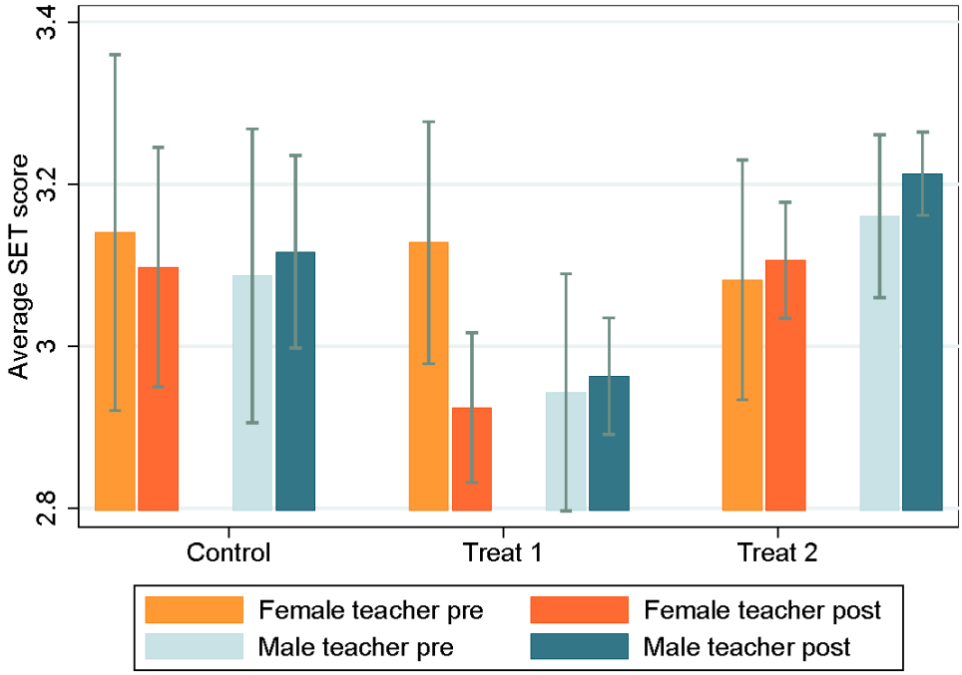


Table 1. Description of the experiment

Group	E-mail type	Number of students	Number of evaluations				
			All	Before email	After mail	After mail, treated	After mail, spillover
Control	None	256	654	205	449		
Treatment one	Normative	525	1,509	315	1,194	617	577
Treatment two	Informative	788	2,329	518	1,811	906	905

Table 2. Descriptive statistics students

	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 specific procedure	.10	.31
Share of students who took the entry exam	.46	.50
Share of students from international procedure	.32	.47
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
Overall satisfaction scores in history	3.21	.82
Overall satisfaction scores in law	3.09	.93
Overall satisfaction scores in micro	3.08	.91

Table 3. Balancing checks

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
	Student female		Note final grade		Note continuous assessment		Age		French citizenship		Entry exam waived		Entry exam		International procedure	
Professor	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men
post	0.038	-0.004	-5.253	-3.832	-2.592	-5.478***	0.276**	0.252***	-0.060	-0.017	-0.036	-0.005	0.0221	-0.035	0.000617	0.0331
	(0.064)	(0.056)	(4.092)	(3.326)	(3.176)	(2.031)	(0.119)	(0.097)	(0.062)	(0.0576)	(0.023)	(0.016)	(0.0550)	(0.047)	(0.0577)	(0.0506)
post*T1	-0.112	-0.087	1.817	-2.009	-2.249	0.475	-0.087	-0.123	-0.0375	-0.160**	0.0274	0.006	-0.184**	-0.191***	0.0875	0.0815
	(0.080)	(0.069)	(4.986)	(4.143)	(3.644)	(2.811)	(0.140)	(0.112)	(0.078)	(0.070)	(0.028)	(0.025)	(0.0722)	(0.063)	(0.0740)	(0.0654)
post*T2	-0.006	0.023	2.182	2.912	-3.141	2.920	-0.126	-0.126	-0.0232	-0.056	0.087**	-0.003	-0.114*	0.031	0.0179	-0.0103
	(0.077)	(0.064)	(4.956)	(3.907)	(3.654)	(2.342)	(0.127)	(0.102)	(0.0649)	(0.059)	(0.038)	(0.028)	(0.0655)	(0.055)	(0.0600)	(0.0519)
Obs	1,733	2,763	1,727	2,746	1,733	2,763	1,733	2,763	1,733	2,763	1,733	2,763	1,733	2,763	1,733	2,763
pval T1 T2	0.101	0.030	0.927	0.126	0.726	0.281	0.647	0.965	0.772	0.0123	0.073	0.785	0.239	1.04e-05	0.158	0.0330

Table 4. Main effects, fall semester courses

	(1) Women	(2) Men	(3) Women	(4) Men	(5) Women	(6) Men	(7) All
Post	-0.079	0.016	-0.077	0.021	-0.078	0.021	0.025
	(0.090)	(0.078)	(0.090)	(0.076)	(0.090)	(0.076)	(0.077)
post*TC1			0.20	0.070			
			(0.14)	(0.11)			
post*TT1	0.091	0.17	0.083	0.17			
	(0.13)	(0.13)	(0.12)	(0.12)			
post*TC2			0.36***	0.018			
			(0.13)	(0.096)			
post*TT2	0.26**	0.054	0.27**	0.053			
	(0.13)	(0.099)	(0.13)	(0.098)			
post*T1					0.14	0.10	0.10
					(0.11)	(0.097)	(0.098)
post*T2					0.31***	0.035	0.033
					(0.11)	(0.087)	(0.088)
post*female							-0.11
							(0.12)
post*female*T1							0.027
							(0.15)
post*female*T2							0.28**
							(0.14)
Observations	1,025	1,542	1,725	2,745	1,725	2,745	4,470
pval T1 T2	0.19	0.33			0.067	0.37	
pval TC1 TT1			0.41	0.40			
pval TC2 TT2			0.47	0.68			
pval TT1 TT2			0.14	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).

Table 5. Robustness checks

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Dummy			Ordered logit regressions			Fall and spring smester Student fixed effects			Fall and spring smester Teacher and student fixed effects		
	Women	Men	All	Women	Men	All	Women	Men	All	Women	Men	All
post	-0.036	0.0069	0.0087	-0.21	0.0027	0.012	-0.28*	0.055	-0.048	-0.19*	-0.046	-0.067
	(0.039)	(0.036)	(0.036)	(0.29)	(0.23)	(0.23)	(0.16)	(0.13)	(0.12)	(0.11)	(0.11)	(0.090)
post*T1	0.055	0.036	0.042	0.37	0.34	0.34	-0.067	-0.19	-0.034	0.24	0.061	0.14
	(0.053)	(0.046)	(0.046)	(0.35)	(0.29)	(0.29)	(0.22)	(0.17)	(0.15)	(0.15)	(0.14)	(0.11)
post*T2	0.15***	0.016	0.014	0.87**	0.10	0.094	0.42**	-0.084	0.029	0.33**	0.022	0.020
	(0.051)	(0.041)	(0.041)	(0.34)	(0.26)	(0.26)	(0.20)	(0.15)	(0.14)	(0.14)	(0.13)	(0.10)
post*female			-0.048			-0.26			-0.17			-0.065
			(0.053)			(0.35)			(0.14)			(0.11)
post*female*T1			0.0053			0.027			-0.058			0.0035
			(0.071)			(0.44)			(0.19)			(0.15)
post*female*T2			0.14**			0.82**			0.36**			0.29**
			(0.065)			(0.41)			(0.17)			(0.14)
Observations	1,727	2,746	4,473	1,727	2,746	4,473	3,465	5,190	8,655	3,202	5,100	8,630
pval T1 T2	0.050	0.55		0.065	0.27		0.0092	0.44		0.50	0.68	

Table 6. Robustness checks, using one treated campus at a time with both control campuses

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Menton		Reims		Poitiers		Paris		Le Havre	
	Women	Men	Women	Men	Women	Men	Women	Men	Women	Men
post	-0.042	0.017	-0.077	0.015	-0.067	0.0069	-0.074	0.033	-0.071	0.029
	(0.092)	(0.075)	(0.089)	(0.074)	(0.090)	(0.073)	(0.090)	(0.076)	(0.090)	(0.076)
Post*T1	0.33	-0.057	0.061	0.10	0.33**	0.26*				
	(0.45)	(0.21)	(0.12)	(0.10)	(0.14)	(0.16)				
Post*T2							0.26**	0.045	0.66***	-0.013
							(0.11)	(0.088)	(0.22)	(0.13)
Observations	397	543	700	922	405	498	991	1,593	422	618

Table 7: Effect of the treatment by student and teacher gender

	(1)	(2)	(3)	(4)	(5)	(6)
Student	Male	Female	Male	Female	Male	Female
Teacher	Female	Female	Male	Male	All	All
post	-0.11	-0.046	0.0051	0.021	-0.34	0.14
	(0.14)	(0.13)	(0.11)	(0.11)	(0.28)	(0.21)
post*T1	0.40**	-0.016	0.24	0.074	0.36	-0.024
	(0.18)	(0.16)	(0.16)	(0.13)	(0.36)	(0.24)
post*T2	0.47***	0.18	0.045	0.041	0.011	-0.12
	(0.17)	(0.15)	(0.13)	(0.12)	(0.31)	(0.24)
post*female					-0.13	-0.0087
					(0.18)	(0.16)
post*female*T1					0.10	-0.0055
					(0.29)	(0.20)
post*female*T2					0.41*	0.12
					(0.23)	(0.20)
Observations	711	1,016	1,055	1,691	1,746	2,676
pval T1 T2	0.68	0.11	0.15	0.71		
pval male vs female student T1	0.070		0.40			
pval male vs female student T2	0.19		0.98			

Note: Fall semester only. All regressions include control variables and teacher fixed effects.

Table 8. 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) Quality of animation	(6) Ability to encourage group work	(7) Availability and communication skills	(8) Ability to relate to current issues	(9) Contribution to intellectual development	(10) Investment	(11) Number of grade	(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	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

Table 9. Heterogeneity of the effect

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Teacher's quality		Student's level		Field		
	> median	< median	> median	< median	Law	Economics	History
post	0.028	0.030	0.067	-0.033	0.011	0.086	0.011
	(0.12)	(0.100)	(0.12)	(0.099)	(0.13)	(0.18)	(0.11)
post*T1	0.19	0.076	0.013	0.21	0.065	0.046	0.17
	(0.17)	(0.12)	(0.14)	(0.13)	(0.16)	(0.22)	(0.17)
post*T2	-0.0016	0.066	-0.030	0.081	0.091	-0.080	0.038
	(0.13)	(0.12)	(0.13)	(0.12)	(0.15)	(0.20)	(0.13)
post*female	-0.14	-0.066	-0.011	-0.24	-0.092	-0.20	0.13
	(0.15)	(0.23)	(0.16)	(0.17)	(0.21)	(0.21)	(0.31)
post*female*T1	0.083	-0.11	-0.11	0.17	0.13	0.11	-0.31
	(0.21)	(0.26)	(0.21)	(0.22)	(0.28)	(0.27)	(0.35)
post*female*T2	0.32*	0.21	0.21	0.33	0.18	0.27	0.26
	(0.18)	(0.25)	(0.19)	(0.21)	(0.25)	(0.24)	(0.34)
Observations	2,154	2,319	2,369	2,104	1,518	1,487	1,468

Note: a teacher is defined as good when the average grade of his students at the final exam is above the average grade in the campus. A student is defined as good when his grade is above average grade in the campus.

Appendix 1. 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),

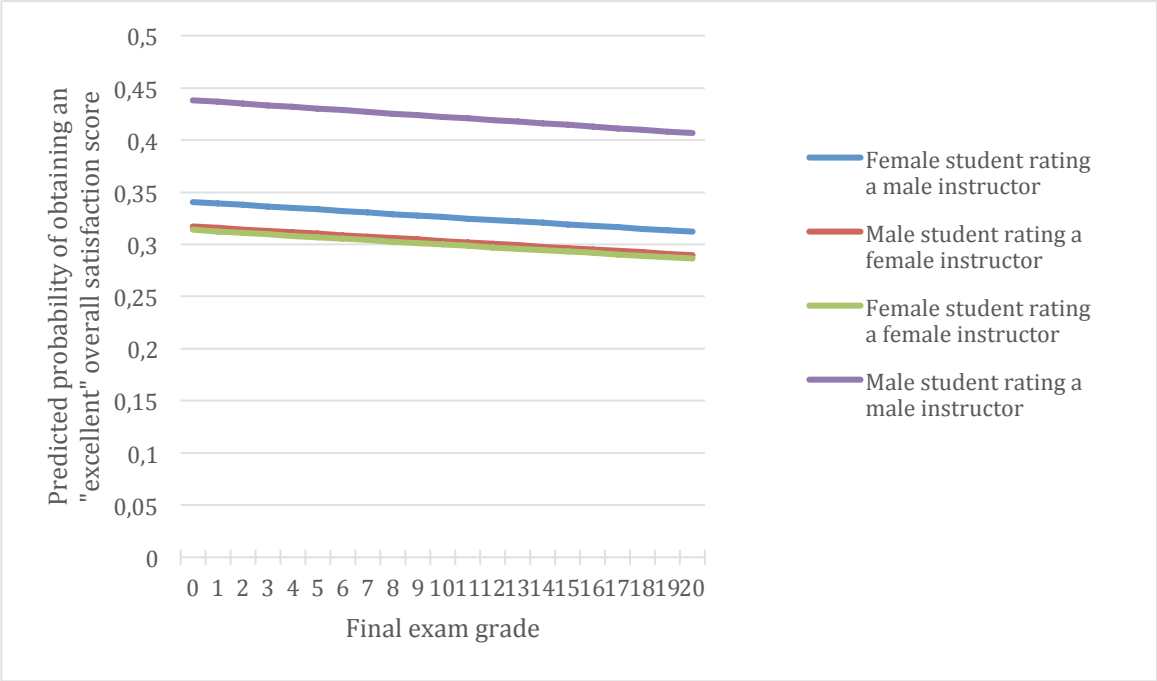
En cette période d'évaluation des enseignements nous souhaitons attirer votre attention sur les résultats d'une recherche récente menée à Sciences Po mettant en évidence un biais discriminatoire à l'encontre des femmes enseignant les conférences de méthode pour les modules fondamentaux de première année.

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.

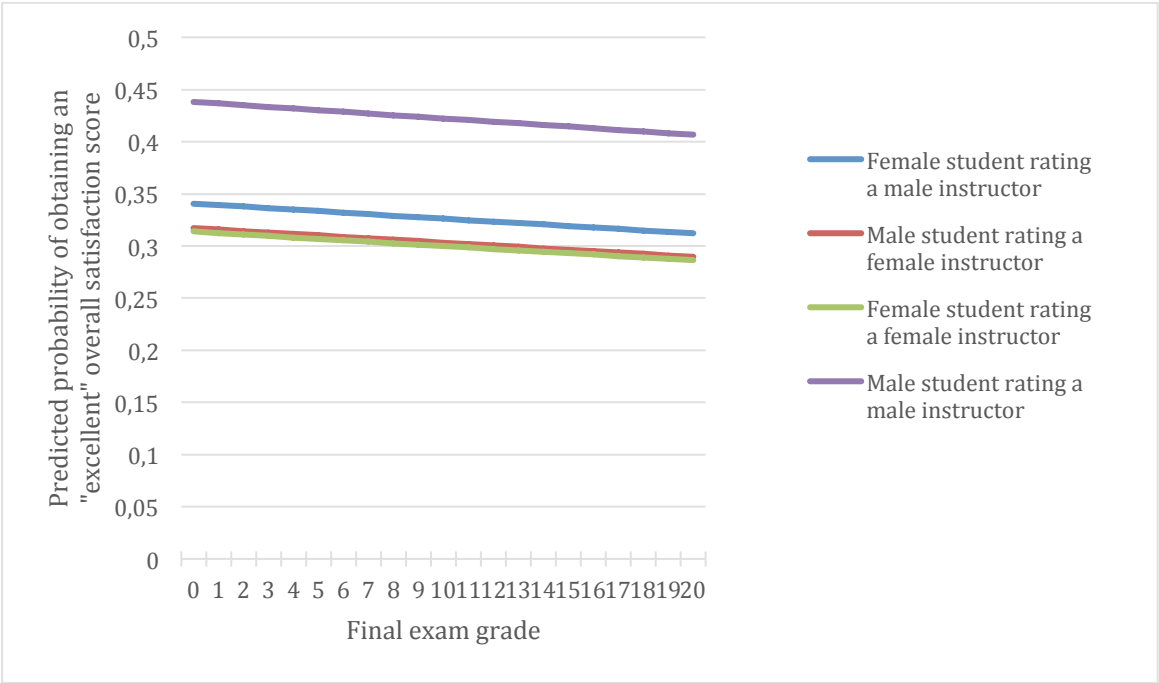
Indeed, the results of this study show that students tend to give lower ratings to their female instructors despite the fact that students perform equally well on final exams, whether their seminar instructor was a man or a woman. Male students in particular tend to rate male instructors higher in their 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 2. 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