I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17632

Gender Bias in Student Evaluations of Teaching: Do Debiasing Campaigns Work?

Sara Ayllón
Camila Zamora

JANUARY 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17632

Gender Bias in Student Evaluations of Teaching: Do Debiasing Campaigns Work?

Sara Ayllón
IZA and EQUALITAS, University of Girona

Camila Zamora
Universitat Autònoma de Barcelona

JANUARY 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Gender Bias in Student Evaluations of Teaching: Do Debiasing Campaigns Work?*

This paper presents the results of a field experiment aimed at reducing the gender bias in teaching evaluations at a higher education institution. In the intervention, before they completed the teaching evaluation questionnaire, students were individually randomized in three groups. One third received an email invitation to watch a video that informed them of the existence of implicit bias (treatment 1). Another third of the students received an email invitation to watch a video with an explicit message that made them aware of the presence of gender bias in teaching evaluations (treatment 2). This second video also mentioned the fact that the academic literature has shown that this form of discrimination often originates with male students. At the end of both videos, all the students treated were asked to avoid displaying prejudice when they completed the questionnaire. The final third of students was assigned to the control group and did not receive any message. The results indicate that the video on implicit bias served to reduce the score gap between male and female lecturers. However, the video on gender bias had an unintended consequence: male students in the treatment group awarded their female teachers even lower scores than did the control group — confirming the risk of backlash or reactance in this kind of debiasing campaign. Such an effect was found to be particularly strong in female-dominated academic contexts.

JEL Classification: C93, I23, J71

Keywords: gender bias, field experiment, gender discrimination, teaching evaluations, higher education, Spain

Corresponding author:
Sara Ayllón
C/Universitat de Girona, 10
17003 Girona
Spain
E-mail: sara.ayllon@udg.edu

* We are extremely grateful for the support received from Natalia Adell, Pepus Daunis, Sílvia Llach and Anna M. Pla Boix at the University of Girona who made implementation of this field experiment possible. We benefited greatly from advice from Olga Stoddard at Brigham Young University during the design of the experiment. Ayllón acknowledges support from the projects PID2022-137352NB-C41 (funded by MCIN/AEI/10.13039/501100011033), 2021-SGR-570 and PATHS2INCLUDE (Horizon Europe, European Commission, grant agrement no. 101094626). Zamora acknowledges support from the scholarship FI SDUR, grant number 00379 (2022). Any errors or misinterpretations are our own.

1 Introduction

Women remain underrepresented in academia and in most STEM fields, despite persistent efforts to narrow the gender gap across various industries and professions. In Europe, for example, the European Commission (2021) reveals that women represent only 41.3% of employed scientists and engineers, even though there are as many women as men in the share of the population with tertiary education; female researchers tend to work more on part-time contracts than their male counterparts; and female academics are underrepresented in full professorial positions (26.2%) and in higher education leadership posts (23.6%). Colby (2023) provides support for these figures in the US, too, where women are more likely to hold non-tenure-track lecturer and instructor positions and are more often on temporary contracts. Despite an increase in women's participation at the bachelor, master and doctoral levels, their careers continue to fall short when compared to men's, particularly in the highest academic positions and in STEM fields (Lincoln et al., 2012; Bayer and Rouse, 2016; McElroy, 2016; Silver et al., 2017; Bagues et al., 2023).

One factor potentially contributing to the persistent gender gap in academia is probably related to the gender bias present in students' evaluations of teaching. Multiple studies have documented the presence of bias against female teachers in such evaluations (Boring et al., 2016; Boring, 2017; Wagner et al., 2016; Mitchell and Martin, 2018; Keng, 2020; Ceci et al., 2023).[1] Female professors regularly obtain worse ratings than their male counterparts, though the gap cannot be explained by differences in teacher effectiveness, performance or skills (MacNell et al., 2015; Mengel et al., 2019). Such results have been found in different contexts (Fan et al., 2019) and teaching modalities (MacNell et al., 2015; Mitchell and Martin, 2018; Ayllón, 2022). Moreover, there seems to be a certain consensus that much of the discrimination arises from male students (Boring, 2017; Mengel et al., 2019; Ayllón, 2022). As a result, women's career advancement and promotion may be jeopardized.[2]

Gender differences in teaching evaluations that cannot be attributed to performance quality must be a result of prejudice or dislike, either conscious or not, either implicit or not (Bertrand et al., 2005; Rooth, 2010; Oreopoulos, 2011; Bohnet, 2016). For example, gender bias may arise because of a discrepancy between the expectations that a student has about a given lecturer and the actual behaviour in class of that lecturer. According to MacNell et al. (2015), students expect male professors to be effective (professional, objective, authoritative and knowledgeable), while female professors are expected to be interpersonal (warm, accessible, nurturing, supportive and personable). Students penalize those instructors that do not conform to their expectations (MacNell et al., 2015; Sinclair and Kunda, 2000; Adams et al., 2022). Moreover, students base their evaluations more on personality when they rate females than when they rate males (Mitchell and Martin, 2018).

To reduce gender bias in teaching evaluations, and in order to correct potential dis-

[1] In this paper, we use 'teacher', 'professor', 'instructor' and 'lecturer' without distinction and regardless of the category or type of contract held.

[2] A few recent studies, however, have found no evidence of bias against female instructors. For example, Andersson et al. (2024) in Sweden, Binderkrantz et al. (2022) in Denmark and Acosta-Soto et al. (2022) in Mexico. In the first two cases, the conclusions arise from experiments in two countries well known for gender equality, which could help to explain why gender-stereotypical expectations are less pronounced and highlight the relevance of context in this kind of analysis. Moreover, the experiments conducted rely on online courses, implying a weaker relationship between the students and the professors they are evaluating. In the Mexican case, the results derive from the evaluation of several teaching competencies.

criminatory practices that can derive from it, a number of interventions have been implemented in higher education institutions in recent years. One group has focused on raising bias awareness among students — in most cases by adding a debiasing message at the beginning of the teaching evaluation questionnaire; meanwhile a second group has implemented other strategies, such as self-affirmation campaigns or interventions that could mitigate likability bias against female professors.[3]

In the case of the first group, Peterson et al. (2019) were — to the best of our knowledge — the first to show that a short introductory text added at the beginning of the teaching questionnaire could change behaviour among students at a university in the US. The text largely explained that the opinions of students influence the annual review of instructors; that the university was aware that evaluation could be influenced by unconscious and unintentional bias; and that women and instructors of colour were systematically rated lower. Finally, students were asked to make an effort to resist stereotyping their professors. The treatment had a significantly positive effect on the evaluation of female lecturers. Also in the US, but with a larger sample, Genetin et al. (2022) reproduced the experiment conducted by Peterson et al. (2019) using another Randomized Control Trial (RCT) that introduced two other messages. In the first, students were given the same message on implicit bias as in Peterson et al. (2019); in the second, they were reminded of the importance for instructors of their evaluation; and in the third, they were shown both warnings. A control group did not receive any message. The results show that, compared to the control group, only the second treatment helped improve the average score of women if they belonged to a racial or ethnic minority — somehow failing to replicate the findings of Peterson et al. (2019). In France, Boring and Philippe (2021) ran a field experiment with three treatment arms: the first group of students received a normative message, consisting of an email encouraging them to avoid discrimination when completing the teaching questionnaire, with a special focus on gender discrimination; the second group received a more explicit message explaining that students, and particularly male students, had previously discriminated against female lecturers at the university in question; and the third group was the control group. The results indicated that (compared to the control group) only the second treatment was effective at reducing discrimination — that is, when students were presented with evidence of gender bias at their institution. Taken together, these first studies indicate that the results of similar interventions seem to depend heavily on the academic context in which they are implemented.

Turning to the second group of interventions, Hoorens et al. (2021), for example, find that a self-affirmation campaign prior to having the students complete the questionnaire (either through a value-affirmation task or self-superiority priming) helped reduce the gender bias in teaching evaluations at a Belgian university. Fisk et al. (2020) focused on addressing gender bias by getting female teachers to provide individual additional positive feedback to a group of treated students. The intervention enhanced the likability of female instructors among top-performing students, ultimately improving the teaching evaluation scores they were awarded.

This paper presents the results of a field experiment aimed at reducing gender bias in teaching evaluations at the University of Girona (Spain), where the existence of such bias has previously been documented by Ayllón (2022) and Ayllón et al. (2024). In the intervention, and before completing the teaching evaluation questionnaire, students were individually randomized into three evenly sized groups. One group received an

[3]Self-affirmation refers to the process of contemplating elements that contribute positively to one's self-image.

email invitation to watch a video that informed them of the existence of implicit bias (treatment 1). A second group of students received an email invitation to watch a video with an explicit message making them aware of the presence of gender bias in teaching evaluations (treatment 2). In this second video, students were also informed that the academic literature shows that such discrimination often originates with male students. At the end of both videos, all the students treated were asked to avoid prejudice when completing the questionnaire. The final group of students was assigned to the control group and did not receive any message.

Our debiasing campaign builds on the findings of previous interventions, particularly those based on anti-bias messages received prior to completion of the questionnaire, but adds two new features that considerably enhance our experiment compared to previous ones. First, our RCT covers all students at a Spanish university. This means that our findings are based on a very large sample and refer to all fields of knowledge, which provides a more nuanced understanding of the potential of such interventions in different academic fields. By contrast, Peterson et al. (2019) covered just two introductory courses at Iowa State University; the intervention by Genetin et al. (2022) was implemented in only two colleges at Ohio State University; and Boring and Philippe (2021) only randomized first-year students at a French university that specialized in Social Sciences. Thus, to the best of our knowledge, a university-wide intervention such as the one we present here has never before been attempted. Second, our anti-bias awareness-raising campaign is based on videos, which exploit the additional effectiveness of audio-visual tools on the target group of young university students (Moss-Racusin et al., 2018). As far as we know, this is the first time that videos have been used in an intervention that aims to reduce gender bias in teaching evaluations.

Our main findings show that, compared to the control group, students treated with video 1 (the 'implicit bias video') ended up giving lower scores to male instructors, which consequently narrowed the gender score gap between male and female professors. Gender bias in teaching evaluations has been found to involve overvaluing male, rather than undervaluing female professors, and our results seem to support that contention (Hoorens et al., 2021). Thus, a focused, timely intervention which simply makes students aware of the existence of implicit bias can be effective in reducing gender bias in teaching evaluations. Our results stand in contrast to those of Boring and Philippe (2021) in France, whose intervention with a normative message had no measurable effect. Regarding the impact of treatment 2 (the 'gender bias video'), the effect of that was the opposite of what we had intended: having been made aware of gender bias in teaching evaluations, and having learnt that such bias often originates with male students, those male students actually became more ready to penalize female instructors (compared to the control group). This is an example of the backlash or reactance that is sometimes found in this kind of intervention (Legault et al., 2011; Moss-Racusin et al., 2014). 'Backlash' refers to a strongly negative reaction to something, while 'reactance' refers to the motivation to protect or regain one's personal freedom — with freedom being the belief that one can engage in a particular behaviour (Miron and Brehn, 2006; Mühlberger and Jonas, 2019). This motivation can elicit a tendency on the part of the target group to do the opposite of what they feel they are being asked to do (Hoorens et al., 2021). In other words, if male students felt that the message in treatment 2 was inappropriate or coercive, they may have decided to 'fight back' and resist the influence of the message. Heterogeneous analysis indicates that this result is driven by the minority of male students who study in a female-dominated field — and in particular, in the faculty of Education and Psychology and the faculty of

Nursing.

This paper contributes to three strands of literature. First, it speaks to the literature that studies labour market inequalities that are a result of subjective evaluations in employment settings. Given that teaching evaluations are used to decide issues of hiring, retention and promotion, the context resembles a common principle-agent problem, where an employer (the university administration) relies on information provided by a third party (the students) about its employees (the professors). If subjective evaluations are biased against a particular group, the employer ends up being involved in discriminatory practices, with adverse professional consequences for the group being discriminated against. This has been found to be true, for example, in academia (Card et al., 2019; Hengel, 2022; Eberhardt et al., 2023), in government (Beaman et al., 2009) and in hiring (Goldin and Rouse, 2000).

Second, we contribute to the literature of economics that underlines the importance of considering stereotypes and implicit bias when studying discriminatory behaviour (Bohnet, 2016; Bordalo et al., 2016, 2019; Bohren et al., 2019). For example, in the context of hiring discrimination, Reuben et al. (2014) show that implicit stereotypes not only predict initial bias in beliefs about the potential performance of males and females, but also foster the sub-optimal updating of expectations when additional information is provided by the candidates themselves. In their lab experiment, employers with stronger implicit bias against females were more willing to believe men's overestimated expectations of their future performance. In a real-life hiring situation, Rooth (2010) documents the fact that the more negative employers' automatic attitudes and performance stereotypes are toward immigrant male job applicants, the lower is the probability that they will be invited for interview. In a similar way, Glover et al. (2017) show that workers underperform when supervised by biased managers, and as a result firms set higher standards for minorities when hiring. In the educational context, Avitzour et al. (2020) find that teachers' implicit gender stereotypes (but not explicit ones) and their underestimation of the stereotypical attitudes they hold are both associated with boy-favouring grading behaviour — see also Carlana (2019).

Third, this paper contributes to the literature on the effectiveness of interventions aimed at reducing discrimination (or increasing diversity) in the workplace (Kalev et al., 2006; Lai et al., 2013, 2014, 2016; Moss-Racusin et al., 2014; Carnes et al., 2015; Bohnet, 2016; Kunze and Miller, 2017; Devine et al., 2017; Bertrand and Duflo, 2017; Chang et al., 2019; Goldin and Rouse, 2000; Arslan et al., 2024; Card et al., 2024; Delfino, 2024). We complement previous research (albeit in a different context) that has shown that informing people about their own potential implicit bias, as in our treatment 1, can be an effective strategy to reduce discrimination (Avitzour et al., 2020; Alesina et al., 2024. On the other hand, our findings on backlash or reactance by male students in some female-dominated faculties (following treatment 2) indicate that messages that are too direct, or that may engender shame or blame within a group, can prove counterproductive (Legault et al., 2011; Dobbin and Kalev, 2016, 2018).

The paper is structured as follows. Section 2 provides all the details of the experiment — from the institutional setting to the different treatment arms, as well as the process of implementing the intervention. In Section 3, we provide details of the data and our identification strategy. Section 4 outlines the results of the study by presenting first our main findings, and then moving on to robustness checks and some heterogeneous results. Section 5 discusses potential mechanisms behind our findings. Finally, in Section 6 we provide our concluding remarks.

2 The experiment

This section describes in detail the institutional setting where our field experiment took place, the different treatment arms and the implementation. The American Economic Association has officially recorded this field experiment as an RCT which was sent on 26 April 2024 — prior to the start of the intervention — and that received the identification number AEARCTR-0013491. Additionally, ethical approval was obtained from the Research Ethics and Biosecurity Committee (CEBRUdG) of the University of Girona on 18 March 2024, with identification number CEBRU0002-24. Both supporting documents can be found in Appendix A to this study.

2.1 Institutional setting

The experiment took place at the University of Girona (UdG) in the spring semester of academic year 2023/24. UdG is one of eight public universities in the region of Catalonia (Spain) and has 50 undergraduate and postgraduate programmes in five main fields of knowledge (Arts and Humanities, Sciences, Life Sciences, Social Sciences, and Architecture and Engineering). It is a medium-sized university, with a population of about 15,000 students. This facilitated close collaboration with the administration and authorities involved in the implementation of the intervention that we designed.[4]

At the University of Girona, lecturers are evaluated through a questionnaire administered online via the Moodle platform. Students can fill in the questionnaire in the three weeks or so before their end-of-year exams. The timing is intended to prevent students' final grades from influencing their responses. In our case, the questionnaire was open from 29 April to 3 July 2024 — the period during which the questionnaire could be completed. The precise timing varied according to faculty and course, and depended on when the end-of-year exams were scheduled; however, the vast majority of questionnaires were active in the first three weeks of that window. The dean of each of the nine faculties determines how to promote completion of the questionnaire. Periodic reminders are also sent out by the Planning and Evaluation Office to the deans, informing them of their respective response rates. This serves as a prompt for them to take the necessary action to improve student participation. At the institutional level, there are some promotional activities to encourage students to participate in the process, primarily through social media. Additionally, a banner is displayed each time a student accesses the Moodle platform, serving as a constant reminder to engage in the evaluation, and further promoting participation in the evaluation process.

Other features of the evaluation programme are as follows. First, the individual responses of students are kept confidential from their instructors, who only learn their average score several weeks after the questionnaire has closed. Second, the questionnaire is identical throughout the university: this allows a comparison of scores across fields of study, faculties and programmes. Third, the questionnaire contains eight questions. Questions 1 to 4 focus on the course itself, and students only answer them once, even if they have had more than one teacher on the course. Questions 5 to 8 enquire about students' opinions of their instructors' performance, and they can answer the questions

[4]There are about 3,000 off-campus students in the six *professional schools* affiliated to the University of Girona. Given the importance in the region of tourism and sports, the great majority of the students receive training in one of those sectors. These students, and particularly their lecturers — who all have an external contract with the university — have a profile that differs from on-campus students and lecturers; as suggested below, this would justify a separate analysis, which we undertake in the 'Mechanisms' section.

for each of the lecturers they have had. In our main analysis, we use question 7, which asks about overall satisfaction with the instructor, while in the 'Mechanisms' section we also look at questions 5 and 6 on the methods used by the lecturer and the activities of support.[5] Fourth, completion of the questionnaire is not mandatory for students, which may imply that our final sample is selected. However, it is the scores of those that fill in the questionnaire that are used for the evaluation of each instructor. Fifth, the questionnaire is asked on all courses where a given lecturer teaches at least 1.5 European Credit Transfer and Accumulation System (ECTS) credits.

2.2 Treatments

For our field experiment we produced two videos that were later sent to two individually randomized groups of students.[6] Treatment 1 consisted of an email invitation to watch a one-minute video (the 'implicit bias video'), which explained in very simple, straightforward language that the time to evaluate professors had arrived; that it was important that students perform such an evaluation; and that when they do so, they should focus on the quality of the teaching, rather than on other aspects of the teaching experience. The video ended by explaining that implicit bias exists and asking students to avoid discriminatory behaviour. The text (as translated from Catalan) read as follows:

From today and for the next few weeks, students at the University of Girona are able to complete the teaching evaluation questionnaire. Your participation in the evaluation of teaching is very important, because it helps your professors to improve their teaching.

It is for this reason that we would like to ask you, when completing the questionnaire, to focus exclusively on the quality of the teaching received and the content of the course, and not on other aspects, such as the professor's gender, appearance, age, country of origin or language used.

In this regard, it is important to avoid implicit bias which shows prejudice, normally negative, towards a given group — bias of which we are not conscious and which often arises involuntarily.

We ask you, therefore, to avoid prejudice or discriminatory behaviour when filling in the questionnaire.

Thank you!

Treatment 2 consisted of an email invitation to watch a different video (the 'gender bias video') of similar duration, but with a more direct message that referred explicitly to gender bias and the importance of avoiding it, in order to ensure a fair evaluation of a professor's performance. The video also mentioned that previous literature had found that gender bias adversely affects female lecturers, and that such bias had been shown to have its origin in the evaluations of male students. The text read as follows:

From today and for the next few weeks, students at the University of Girona are able to complete the teaching evaluation questionnaire. Your participation in the evaluation of teaching is very important, because it helps your professors to improve their teaching.

[5] Questions 4 and 8 require students to write down comments if they so desire.

[6] The videos were produced by the authors of this study using the software *Canva* and at no cost. Canva is a graphic design platform that provides tools for creating social media graphics, presentations, promotional merchandise and websites. Both videos will be made available on the personal webpage of the first author of this paper.

It is for this reason that we would like to ask you, when completing the questionnaire, to focus exclusively on the quality of the teaching received and the content of the course, and not on other aspects, such as the professor's gender, appearance, age, country of origin or language used.

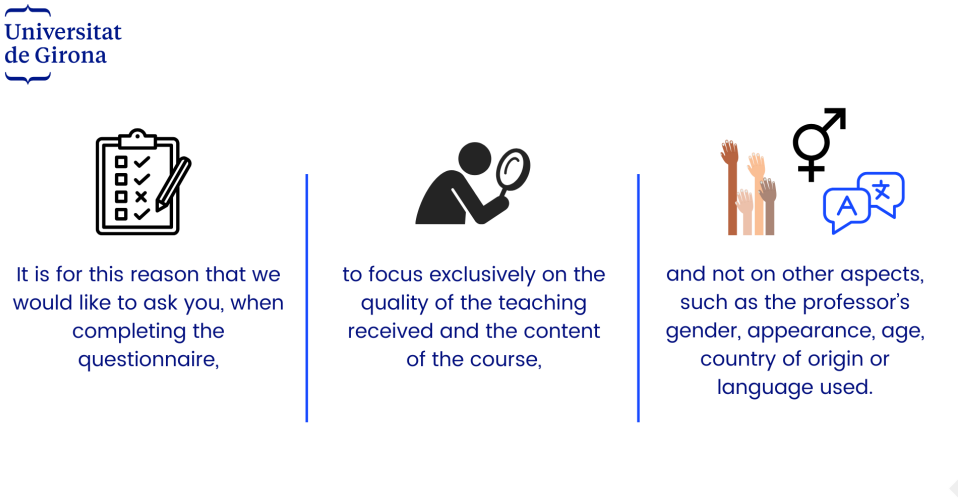
The academic literature has shown that, on average, female professors receive lower scores for their teaching compared to male professors, even when the quality of their teaching is the same. Additionally, it has been documented that such biased behaviour often arises from male students (and not so much from female students).

We ask you, therefore, to avoid prejudice or discriminatory behaviour when filling in the questionnaire.

Thank you!

Thus, while video 1 provided only a general context to make students aware of the existence of bias (irrespective of the source), video 2 specifically aimed to make students aware of bias against female professors, anticipating the potential for biased behaviour during the evaluation process. Other important features of the videos are as follows. First, the videos used a very simple aesthetic, employing the letter type and corporate colours of the University of Girona (mainly white and blue). The text of the message was accompanied by icons or symbols meant to be gender neutral and designed to avoid any favouritism in the aesthetics. To this end, we avoided: (1) the use of videos produced by AI, as they often use explicit images of men and women; and (2) the use of a voice that would read the text, as inevitably it would have had a gendered sound. Figure 1 provides an image of the videos, while the video sequences can be found in Appendix B. Second, the videos were held at the University of Girona YouTube institutional channel, making them unavailable for downloading. Finally, we employed an encouragement design: students were invited to watch the video, but watching it was not mandatory to proceed with the evaluation.

Figure 1: Image of the treatment videos



Note: Authors' elaboration, using the videos produced in the experiment.

2.3 Implementation

All students invited to answer the teaching questionnaire for a course during the second semester of academic year 2023/24 were randomized. In total, there were 14,164 students. The unit of randomization was at the individual (student) level, stratified by student gender to ensure balance. Thus, one third of the students were assigned to treatment 1 (4,713 students), one third to treatment 2 (4,720 students) and one third to the control group (4,731 students).

The day that the evaluation period started, treated students received an email from the vice-rector for Quality and Transparency, explaining to them that from that moment on, they could fill in the teaching evaluation questionnaire; they were also asked to watch a video, available by clicking on a link to the university's YouTube channel. The message was purposely sent out on the very first day that the questionnaires became active. The message from the vice-rector was the same for both treated groups and read as follows:

Dear Student,
From today and for the next few weeks, students at the University of Girona are able to complete a questionnaire that evaluates professors. In this regard, we would like to ask you to watch a video at the following link, before responding to the survey for this semester's courses: [video treatment 1 – youtube link] or [video treatment 2 – youtube link]
Thank you in advance,
Best regards,
Vice-Rector for Quality and Transparency

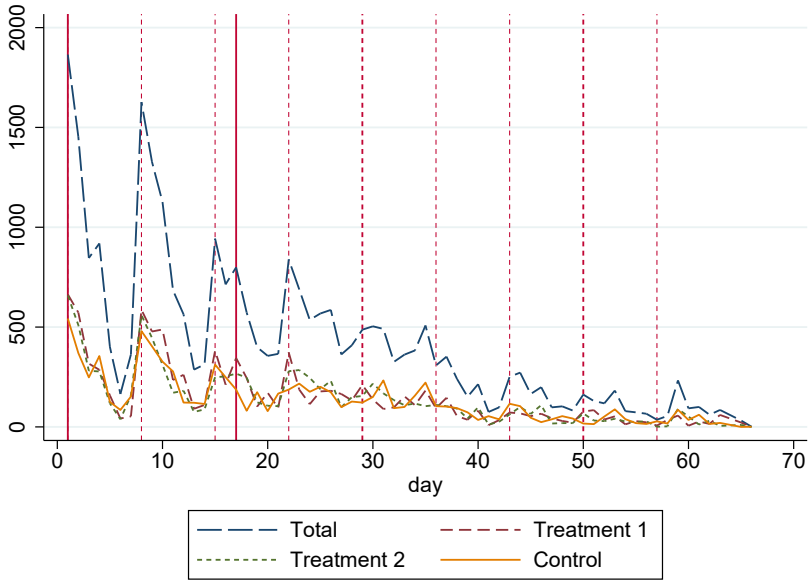
Some two weeks after the start date, on 15 May 2024, any student who still had to complete a questionnaire received a new message from the vice-rector, with a reminder to fill in the evaluation after watching the video.

Figure 2 shows the total number of questionnaires completed and the number completed, by treatment arm, from the start of the evaluation period (and the field experiment). The solid vertical lines indicate the day on which students received the first message (day 1 of the experiment) and the day on which they received the reminder (day 17 of the experiment). The dashed vertical lines mark all Mondays. Three important features are worth highlighting from the figure. First, the number of questionnaires completed: the graph indicates that the vast majority of questionnaires were completed in the first three weeks of the evaluation period — with Monday being the day of the week on which most were filled in. Second, the greatest number of questionnaires were completed during the second week of the experiment, while the reminder email had a fairly small effect: there is little indication that it prompted more responses. Third, the figure indicates that in the first three weeks of the experiment, treated students (particularly those in treatment 1) provided a slightly greater number of responses to the questionnaire than students in the control group.[7] This is confirmed by the fact that, of the 4,713 students assigned to treatment 1, 1,549 filled in at least one questionnaire; of the 4,720 students assigned to treatment 2, 1,518 completed at least one questionnaire; while of the 4,731 students in the control group, 1,460 did so. This is a response rate of 32.8%, 32.0% and 30.8%, respectively, indicating that the messages may have helped increase the response rate, but not to any great extent.

As for the number of times the videos were watched (information that we obtained

[7]This pattern is less clear from the third week of the evaluation period.

Figure 2: Number of responses to the teaching evaluation questionnaire by day, total and by treatment arm, University of Girona, second semester of academic year 2023/24



Note: Authors' elaboration, using data from the University of Girona.

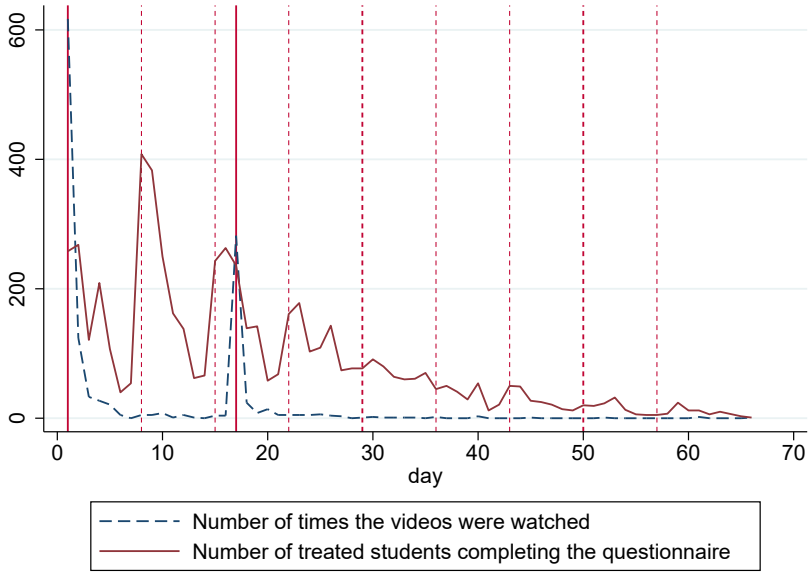
from YouTube), not all treated students complied with what they were asked to do: while a total of 9,433 students were treated, in all, the videos were watched just 1,238 times. That is a rate of 13.1%, slightly below the rate of 15% in Arslan et al. (2024) for an intervention that asked professional hiring managers in a multinational company to watch a diversity training video. However, if instead of focusing on the total number of treated students, we consider those who went on to complete the questionnaire, the rate is larger. If we were to assume that all treated students who proceeded to answer the questionnaire had previously watched the video, the rate would be 40% (1,238 of the 3,067 treated students who effectively completed the questionnaire). While we cannot be sure that 100% of the treated students who answered the questionnaire first watched the video that they were assigned, it is likely that a large number of them did so, because these tend to be the students who generally comply with what they are asked to do.[8] The dashed line in Figure 3 indicates that most students watched the videos on the first day of the experiment and on the day when they received a reminder: on just those two days, the number of viewings was larger than the number of treated students who completed at least one questionnaire (solid line).[9] As for the other days, the number of viewings was very low, close to zero on most days — possibly indicating that spillovers from treated individuals to the control group were limited. Both treatment videos were watched a similar number of times. As a result, the intervention captures an intention-to-treat effect which is necessarily a lower bound estimate of the potential impact, as only a fraction of

[8]Naturally, we cannot know whether a student watched the assigned video more than once, but we believe this to be unlikely.

[9]About 80.3% of students completed all the questionnaires in a single day; 15.8% took two days; and only 3.9% took more than two days. This means that when the intervention impacted the student, it impacted all his/her responses to the questionnaires. On average, students completed eight instructor questionnaires.

the students treated ended up viewing the videos.

Figure 3: Number of times the videos were watched and number of treated students completing at least one questionnaire per day



Note: Authors' elaboration, using data from the University of Girona and YouTube.

3 Data and identification strategy

3.1 Data

Our main empirical analysis is based on the responses of on-campus students to the teaching evaluation questionnaire in the spring semester of academic year 2023/24, following implementation of the anti-bias intervention. The data, completely anonymized, was provided to us by the IT services of the University of Girona. The overall number of responses to the teaching questionnaire was 27,257. In all, 3,223 on-campus students rated 1,439 instructors on 1,062 different courses across all disciplines of study.[10]

Table 1 presents the summary statistics for our data. Panel A shows that students have an average age of 22 years. About 62% of the responses were from female students. The largest share of answers was from students aiming to obtain a degree in Social Sciences (29% of the sample), followed by Life Sciences (20%). Most of the students who completed the questionnaire were in their first year (about 39%), while those in their fourth and fifth years represented a very small part of the sample (9%). Regarding the characteristics of lecturers (Panel B), the data indicates that the average age of the professors evaluated was 49. About 48% of the teaching evaluation responses referred to a female lecturer. The great majority of professors being evaluated were on a temporary contract: only 39%

[10]Note that there are more teachers being evaluated than there are courses, as it is not unusual for teachers to share a course. We were provided with 7,000 additional scores from 1,304 off-campus students at the university's professional schools, who rated 226 instructors on 318 courses. See the corresponding analysis in the 'Mechanisms' section.

of responses referred to a lecturer on a permanent contract. As a matter of fact, 53% of the responses evaluated a non-faculty member (neither a full, associate or assistant professor).

Table 1: Summary statistics

	Mean	Std. Dev.	Min.	Max.
Panel A: Student characteristics				
Age	22.23	4.89	18	68
Female	0.62	0.48	0	1
Arts and Humanities	0.05	0.21	0	1
Sciences	0.21	0.41	0	1
Life Sciences	0.20	0.40	0	1
Social Sciences	0.29	0.45	0	1
Architecture and Engineering	0.24	0.43	0	1
1st year	0.39	0.49	0	1
2nd year	0.32	0.46	0	1
3rd year	0.21	0.41	0	1
4th and 5th year	0.09	0.28	0	1
Students' grade	7.16	1.63	0	10
Panel B: Lecturer characteristics				
Age	48.63	10.30	23	74
Female	0.48	0.50	0	1
Permanent contract professor	0.39	0.49	0	1
Full professor or emeritus	0.07	0.26	0	1
Associate professor	0.32	0.46	0	1
Assistant or visiting professor	0.08	0.28	0	1
Adjunct or pre-PhD faculty	0.53	0.50	0	1
Average score	3.98	1.19	1	5
% received a score of 1	0.06	0.23	0	1
% received a score of 2	0.07	0.25	0	1
% received a score of 3	0.16	0.37	0	1
% received a score of 4	0.26	0.44	0	1
% received a score of 5	0.45	0.50	0	1

Note: Each observation is at the professor-student-course level.
Source: Authors' computation, using data from the University of Girona.

Our main dependent variable contains the responses to the statement in the questionnaire that asks students for their overall assessment of the lecturer's work: 'I evaluate this teacher's overall performance as positive.' Responses are on a Likert scale ranging from 1 to 5, where 1 indicates 'strong disagreement' and 5 'strong agreement'. In general, students are fairly satisfied with the teaching of their lecturers at the university: the average score is 3.98. The last rows of Table 1 also indicate that nearly half of the evaluations awarded the maximum score (45%), while only about 13% gave the lowest score possible (1 or 2).

3.2 Identification strategy

We use a linear regression model with fixed effects to evaluate the intervention and quantify the causal impact of the two treatments on students' responses on the teaching evaluation questionnaire. Formally, we estimate the following equation:

$$Y_{ijs} = \alpha + \beta_1 \text{Treatment } 1_i + \beta_2 \text{Treatment } 2_i + \gamma X_i + \theta_j + \lambda_s + \epsilon_{ijs} \quad (1)$$

where Y_{ijs} is the overall satisfaction score that student i gives to professor j on course s. Treatment 1 (treatment 2) is an indicator variable that takes value 1 if student i was assigned to the first (second) video and 0 otherwise. X_i is a vector of student-level characteristics, in particular age, age-squared, gender (binary) and final grade obtained.[11] θ_j are professor fixed effects, λ_s are course fixed effects and ϵ_{ijs} is the error term. Professor fixed effects control for time-invariant characteristics of professors that might influence the teaching evaluation scores, such as teaching style, performance quality or personality, while course fixed effects control, for example, for the level of difficulty of a given subject or the workload, which can influence the evaluation (Ayllón et al., 2019). By including fixed effects by teacher and course, we can hold constant all shared aspects of the learning experience.[12] The inclusion of professor- and course-level fixed effects implies that our results identify from the variability of scores provided by (randomly assigned) treated and control students evaluating the same lecturer teaching a given subject. This way, we guarantee that our results are not affected by teaching performance and teacher quality; still less are they affected by the possibility that lecturers provide different quality when teaching different courses.[13]

Our coefficients of interest are β_1 and β_2 which indicate how the treatment assignment affects the evaluations of teaching, after controlling for student characteristics, as well as teacher- and course-level fixed effects.[14] Given that, as explained above, not all treated students watched the video they were sent, these are intent-to-treat effects. In any case, these are usually the results of interest for organizations and institutions — on the assumption that, in the great majority of cases, they have control over treatment assignment, but not take-up. To further account for potential heterogeneity in the treatment effects, we stratify the analysis by professor gender, estimating separate regressions for male and female lecturers. We also investigate the potential differences in treatment effects between male and female students, allowing us to capture any potential interactions between them and the professor's gender. Robust standard errors are used throughout the analysis.

Finally, Table C.1 in Appendix C presents mean differences across treatment arms prior to the intervention, in order to check whether the randomization process was successful in distributing the intervention evenly — even though we only stratified by student gender. This is key to our identification strategy. Indeed, with only one exception, differences across treatment arms are not statistically significant at conventional levels (95%). It is important to recall at this point that this is not the final sample of students used in the analysis, since at the University of Girona it is not mandatory to complete the teaching questionnaire. Moreover, the intervention could have altered the pool of students that ended up filling in the questionnaire. This is an aspect that we take up in the 'Mechanisms' section of this study.

[11] As in Boring (2017) and Mengel et al. (2019), we use the final grade obtained on the course to control for teaching quality.

[12] Naturally, in this setting, no researcher can have complete control of confounders — which may be observed by the students, but not by the researchers; however, the inclusion of both teacher- and course-level fixed effects can help mitigate the effect of unobservables.

[13] We would expect a lecturer to be better at teaching a course for which s/he has more experience, for example.

[14] β_1 and β_2 are the causal effects of the treatment assignments under the identifying assumption that treatment assignments are orthogonal to the error term, which holds by design.

4 Main results

Table 2 presents our main findings for the effects of the two treatments on the overall teaching scores of male and female lecturers (Panel A) drawn from Equation (1). Each column is the result of a separate regression. Column (1) presents the results of the impact of the intervention when all students evaluate their male lecturers, while column (2) does the same for female lecturers. Columns (3) to (6) separate the results not only by professor gender, but also by student gender.

Overall results in the first two columns indicate that treatment 1 (the 'implicit bias video') had a negative effect on the teaching scores of male professors (compared to the control group), while it did not have any effect on the evaluation of female professors. As such, this treatment arm helped to reduce the gender score gap in the teaching evaluations. In Ayllón et al. (2024), it is documented that the potential gender bias in previous semesters at UdG stood at 0.046, being statistically significant at 95%.[15] Assuming comparability, this would mean that treatment 1, with a coefficient of -0.069, partially corrects for the existing gender bias, provided only one third of the students received this treatment. At the same time, treatment 2 (the 'gender bias video') brought no change in the behaviour of treated students (compared to the control group), irrespective of whether they were evaluating their male or their female lecturers. However, the results are more nuanced if we also account for student gender. Two important findings are worth highlighting.

First, and as for treatment 1, the results indicate that the negative effect of the video on the evaluation of male lecturers comes from both male and female students. That is, a fresh awareness of the existence of implicit bias caused both groups of students to lower their evaluation of male professors, while it maintained the scores for female lecturers. Thus, as a consequence of the first treatment arm, all students contributed to the reduction in the scores gap. At this point, it is important to remind ourselves that the 'implicit bias video' simply informed students of the existence of unconscious bias and asked them to pay attention to the quality of the teaching received, and not to other aspects of their teaching experience.

Second, treatment 2 had no distinct impact on the scores of male lecturers, irrespective of whether the evaluation was provided by male or female students. Both groups awarded male lecturers slightly lower scores, but none of the coefficients was statistically significant at conventional levels. However, making male students aware of the existence of gender bias (and stressing that such bias could often be laid at their door) had the opposite effect to the one intended: compared to the control group, treated male students awarded lower scores to their female lecturers — yet another example of the risk of backlash or reactance in this kind of intervention (Legault et al., 2011; Moss-Racusin et al., 2014; Bohnet, 2016; Dobbin and Kalev, 2016, 2018). At the same time, female students who received treatment 2 awarded their female lecturers slightly higher scores; however, the impact was fairly small and did not attain statistical significance. Note, however, that this was the only positive coefficient in the table.

All in all, while the 'implicit bias video' (treatment 1) served to increase awareness and reduce the evaluation gap between male and female professors, the 'gender bias video' (treatment 2) exacerbated the situation, since treated male students ended up awarding even lower scores to their female lecturers (compared to the control group). Despite

[15] Note that this is an average: in some contexts there is no gender bias at all, while in others it is much greater.

Table 2: Main results

	All students		Male students		Female students	
	Male lecturer	Female lecturer	Male lecturer	Female lecturer	Male lecturer	Female lecturer
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: By treatment						
Treatment 1	-0.069***	-0.027	-0.068*	-0.037	-0.064**	-0.009
	(0.02)	(0.02)	(0.04)	(0.05)	(0.03)	(0.03)
Treatment 2	-0.020	-0.031	-0.020	-0.128***	-0.020	0.011
	(0.02)	(0.02)	(0.04)	(0.05)	(0.03)	(0.03)
Observations	13281	12492	5641	3582	7324	8636
R2	0.363	0.361	0.407	0.394	0.403	0.396
Panel B: Intervention						
Treated	-0.045**	-0.029	-0.045	-0.081*	-0.042	0.000
	(0.02)	(0.02)	(0.03)	(0.04)	(0.03)	(0.02)
Observations	13281	12492	5641	3582	7324	8636
R2	0.363	0.361	0.407	0.393	0.402	0.396

Note: Each column in each panel is the result of a different regression, where the dependent variable is the teaching evaluation score given to lecturer j by student i relative to course s. All regressions include student characteristics (age, age-squared and final grade obtained) and professor- and course-level fixed effects. Columns (1) and (2) also control for students' gender. The number of observations in columns (3) and (5) may not add up to those of column (1) because the number of singleton observations in each regression may differ. The same is true in the case of columns (4) and (6), in relation to column (2). Robust standard errors in parentheses. *** significant at 1%, ** at 5% and * at 10%.
Source: Authors' computation, using data from the University of Girona.

such behaviour, when evaluating the whole intervention (both treatments together), we find a negative coefficient for male lecturers (statistically significant at 95%), while the coefficient for females is indistinguishable from zero — see Panel B in Table 2. Therefore, we can confirm that the intervention helped to narrow the gap in the scores of male and female instructors, in spite of the unintended effect brought about by some treated male students when they evaluated their female lecturers.

4.1 Robustness checks

Our results are robust to several specifications. In Panel A of Table 3 we present the results of a specification that, instead of using both fixed effects by lecturer and course, we include only fixed effects by lecturer and we control for course-level characteristics. Results are very similar to those presented in Table 2, with a negative coefficient for males in association with treatment 1 and confirmation of a backlash/reactance effect that originates from male students who received treatment 2 in their evaluation of female lecturers. However, the negative coefficient observed for treatment 1 proceeds largely from female students, and less so from male students — in contrast to what we saw in the main results table. Coefficients are similar in Panel B when, in addition to the inclusion of lecturer and course fixed effects, we use robust standard errors clustered at the lecturer

level — which is reassuring.

Following the train of thought in Boring and Philippe (2021), when evaluating a similar intervention, Panels C and D estimate the same type of previous models, but using a dummy as a dependent variable: in the first case, by grouping scores 1 and 2 on the one hand, and scores 3 to 5 on the other; and in the second case, by simply separating non-excellent scores (1 to 4) from excellent scores (5). Results indicate that the reduction in the score gap between male and female professors associated with treatment 1 is mostly present because of a decline in the probability of both male and female students awarding their male lecturers a score of 5. By contrast, the backlash (or reactance) effect affects female lecturers, who are less likely to receive a score of above 2 from male students, although it does not reduce the likelihood of them receiving a score of 5.

Finally, we thought to assess our main findings by computing the results for those lecturer-course combinations that have at least three observations in the control group and at least three observations in one of the treated groups. The results point in the same direction, with a negative coefficient from treatment 1 in the evaluation of male professors (which again implies a narrowing of the gender score gap at the University of Girona) and an even stronger backlash/reactance effect in treatment 2, stemming from the evaluation of female instructors by their male students.

4.2 Heterogeneous effects

Our main results may affect subgroups of professors differently and may arise from different subgroups of students. For instance, it could be that the backlash/reactance we find only comes from poorly performing students; or it could be that it affects only young lecturers. In what follows, we investigate this by taking account of the characteristics of professors and students. This may provide a deeper understanding of the effects of the intervention. All the tables commented on below can be found in Appendix C.

Tables C.2 and C.3 present the results by the age and tenure status of lecturers, respectively. In the first table, if we look at the negative impact of treatment 1 for male lecturers, our findings suggest that the effect mainly derives from male students evaluating male professors who are aged 50 or over, while the backlash/reactance effect observed in male students' evaluation of female lecturers (treatment 2) is only present when the lecturers being evaluated are aged 35 or over.[16] If we look at the tenure status of lecturers (Table C.3), the results tally with those of Table C.2. As for the negative effect of treatment 1 on male professors, the results suggest that this is mainly driven by male students evaluating associate professors. As for the backlash effect, this is largest when male students evaluate female associate professors — usually at prime age in their academic career. The coefficient is also large among assistant professors and adjunct faculty, but only the latter reaches statistical significance at 90%.

Next, we consider separate regressions by two student characteristics. First, we account for the final grade obtained by students on each course. We consider four categories: '1' for those students who ended up failing the course; '2' for students who passed the exam, but did not obtain a good grade (between 5 and 6.9); '3' for students who got a grade of between 7 and 8.9; and finally, '4' for students who achieved a grade of 9 or more. If we consider the results shown in Table C.4 regarding the negative effect on male teachers as a consequence of treatment 1, it is worth noting that this is mainly driven by the

[16]Note that there are few observations of male students evaluating female lecturers under 35 years of age.

behaviour of students who did not perform too well, but simply passed the course. The coefficients also indicate that the backlash/reactance effect was strongest among those with a grade of between 7 and 8.9. Furthermore, we would like to highlight the divergent behaviour of male and female students who obtained the highest possible grades ('excellent'): female students awarded male lecturers a lower score and female lecturers a higher score, regardless of the treatment received; however, male students who received treatment 2 did the opposite, awarding better scores to their male lecturers. However, not all these coefficients attain statistical significance.

Second, Table C.5 looks at students' level of seniority — defined as when their courses are typically undertaken (between the first and the fifth year).[17] The results indicate that the negative impact of treatment 1 on male professors is driven primarily by female students in their first year. Male students treated with the 'gender bias video' are more critical of their female professors in their second, fourth and fifth years. Third-year students demonstrate no significant change in their responses to the intervention. Meanwhile, the first-year cohort does not exhibit any sign of backlash or reactance. The findings suggest that the response of younger students differs from that of older students, implying that, as students advance in their degrees, the evaluation of their lecturers becomes more biased against women.

In summary, the results show that the impact of treatment 1 — which reduced the scores achieved by male professors and helped narrow the gender score gap — is greater in the case of older lecturers (aged 50 or more) and those who are at the associate level of tenure; and that treatment had a greater effect among students who passed the course, but did not obtain a high grade. On the other hand, the most worrying consequence of the intervention comes in the form of a backlash/reactance effect arising from treatment 2, which resulted in some male students awarding their female professors lower scores: this is particularly the case if the lecturers are over the age of 35 and are at associate professor level; and if the students concerned have achieved a good grade, primarily in either their second or their last year.

5 Mechanisms

Several mechanisms could be driving our results. First, given that completing the questionnaire is not mandatory at the University of Girona, it could be that the intervention did change the composition of students who typically fill in the questionnaire. For example, there is a possibility that treatment 2 could have incentivized more female students to complete the questionnaire in an attempt to counterbalance the bias that (they now learn) exists against female teachers. Figure 2 above and Table C.6 in the Appendix indicate that while slightly more treated students completed the questionnaire than in the control group, there was no substantial change in the characteristics of treated students, compared to the control group.[18]

Second, rather than student and professorial characteristics (analysed in the previous section), it could be that our results are driven by the context in which students develop: the different fields of knowledge present at the University of Girona, the actual faculties

[17] These results should be treated with caution, as the number of missing values for this variable is unusually large; it was therefore not used for the main results. Also, one needs to take into account that undergraduate degrees at the University of Girona generally take four years; only a few require five years.

[18] Table C.6 runs Equation (1) with students' characteristics and field of knowledge as dependent variables.

where students spend their days and the extent to which students are in a male- or female-dominated academic context. It is possible that the intervention was understood, accepted and applauded differently in different environments. Tables 4 and 5 present the results of the debiasing campaign on the teaching scores of male and female lecturers by field of knowledge and faculty, respectively. The effect of treatment 1 on the decline in the scores of male professors — and consequently, the narrowing of the gender score gap — shows a mixed pattern. In Life Sciences, results are mostly driven by female students awarding male lecturers lower scores — which actually compensates for the higher scores given by a minority of male students evaluating male instructors. Instead, in Social Sciences, male students are primarily responsible for the negative evaluations of male professors. At faculty level, we find statistically significant negative coefficients in the Economics, the Education and Psychology and the Medicine faculties. Yet, whereas in the first case the results are mostly driven by male students, in the other two cases they mainly stem from the evaluations by female students.

In turn, Table 4 indicates a mild backlash/reactance effect on the part of male students evaluating female professors in Social Sciences, and a strong one in Life Sciences; Table 5 narrows this down to show that the effect is concentrated in the Education and Psychology and the Nursing faculties. In the former faculty, students pursue degrees in both Social Sciences (such as Pedagogy, Teacher Education or Social Work) and Life Sciences (mostly Psychology); in the latter, all students pursue a degree in Life Sciences — namely Nursing. What is the characteristic that both faculties share that triggers such an important backlash/reactance effect? They are strongly female-dominated contexts — at both student and faculty level. This is clear when we compute the percentage for the total number of completed questionnaires from male students over the total number from female students: quite unlike the rest of the faculties, the figure is less than 20% in both contexts. This is an important aspect to take into account when evaluating the intervention: the results underscore the fact that treatment 2 triggers the strongest backlash/reactance effect from male students against female lecturers in those contexts where male students are in the minority.[19]

Additional results in Table C.7 show the null impact of the intervention in the five off-campus professional schools affiliated to the University of Girona and based not in Girona itself, but in other cities or towns. In these schools, students receive training from professional practitioners, who typically have another job and come to the school to teach a course that is closely linked to their employment or career. Treated students did not let the experiment influence their opinion of these professionals at all, irrespective of whether they were men or women.

Finally, in Table C.8 we explore the effect of the treatments on the other two dimensions of a lecturer that students also evaluate. Results indicate that the narrowing of the gender score gap associated with treatment 1 is also present when students rate the methods used by the professor (question 5) and the activities of support (question 6). At the same time, the backlash/reactance effect associated with treatment 2 is also present in both questions — indicating that male students are critical when assessing not only the overall performance of a female lecturer, but also these other aspects.

[19] According to data obtained from the University of Girona, the faculties of Education and Psychology and Nursing are the only ones where the percentage of female lecturers is above 60%. Also, females are 86% of the student body in the faculty of Nursing and 79% in the faculty of Education and Psychology. These are the largest numbers across faculties at this university.

6 Concluding remarks

A number of lessons can be learnt from our university-wide debiasing campaign, where we seek to reduce gender differences in student evaluations of teaching. First, increasing awareness by informing students of the existence of implicit bias through a direct, simple message that also asked them to avoid displaying prejudice was sufficient to induce a change in behaviour that helped narrow the gender score gap. These results align with recent findings by Avitzour et al. (2020) and Alesina et al. (2024) in an educational context. Thus, a way forward for higher education institutions could be to simply make students aware of their own implicit bias, as a way of correcting discriminatory conduct. Exposing current students, who one day will be in positions of power, to their own bias and training them to overcome such prejudice will promote gender equality in future societies. However, our results differ from those obtained from a similar experiment run in France by Boring and Phillipe (2021), whose normative message asking students not to discriminate had no impact. Future research should focus on understanding why relatively similar campaigns in relatively similar educational contexts should yield different results.[20]

Secondly, we learn that students can be very sensitive to the message conveyed: they may feel shame, blame or anger and, as a result, may react in a way that runs counter to what is intended. This is what likely happened among male students in treatment 2 (the 'gender bias video') in two female-dominated contexts at the University of Girona — the Education and Psychology and the Nursing faculties. Finding out that the literature has identified that discrimination towards female lecturers often originates with male students can lead to an even more negative evaluation by male students of their female instructors. Warning students against gender bias can exacerbate the very gender bias that it seeks to eradicate. It could be that the message renders the professor's gender even more salient in the evaluation. It could also be that the message is felt to be inappropriate or coercive, thus eliciting a backlash or reactance (Mühlberger and Jonas, 2019).

Thirdly, future interventions should consider ways of doing more to reach young students at higher education institutions, as not all treated students followed the instructions received by email from the vice-rector, one of the most important figures in the administration of the university. Given the large number of messages and social media content that they receive on a daily basis, students should probably be further incentivized. Because of the relatively low rate of compliance, our results are necessarily a lower bound of what an intervention could achieve. Equally importantly, universities need to find ways of incentivizing the rate of response to the questionnaires, if they plan to keep using these as a tool for the evaluation of professors (Neckermann et al., 2022).

Much further work needs to be done if we are to fully understand the type of messages or tools that young university students need to receive, in order to maximize the effects of interventions that aim to reduce gender bias in academia — and ultimately, to reduce gender inequality. The few existing experiments in this field so far do not reach any consensus and appear to provide context-dependent results. Nailing the type of intervention necessary to reduce gender discrimination without eliciting a backlash, reactance or overcorrection is not an easy task. Further research on this is warranted.

[20] An important difference could be that completion of the questionnaire was mandatory in the case of Boring and Philippe (2021), whereas in our case it was not.

References

Acosta-Soto, L., K. Okoye, C. Camacho-Zuñiga, J. Escamilla, and S. Hosseini (2022): "An analysis of the students' evaluation of professors' competencies in the light of professors' gender," paper presented at the 2022 IEEE Frontiers in Education Conference.

Adams, S., S. Bekker, Y. Fan, T. Gordon, L.-J. Shepherd, E. Slavich, and D. Waters (2022): "Gender bias in student evaluations of teaching: Punish[ing] those who fail to do their gender right," *Higher Education*, 83, 787–807.

Alesina, A., M. Carlana, E. La Ferrara, and P. Pinotti (2024): "Revealing stereotypes: Evidence from immigrants in schools," *American Economic Review*, 114(7), 1916–1948.

Andersson, O., M. Backman, N. Bengtsson, and P. Engström (2024): "Are economics students biased against female teachers? Evidence from a randomized, double-blind natural field experiment," SSRN 5020176.

Arslan, C., I. Bohnet, E. Chang, S. Chilazi, and O. Hauser (2024): "Just-in-time diversity training leads to more diverse hiring in a global engineering firm," Unpublished working paper.

Avitzour, E., A. Choen, D. Joel, and V. Lavy (2020): "On the origins of gender-biased behavior: The role of explicit and implicit stereotypes," Working Paper 27818, National Bureau of Economic Research.

Ayllón, S. (2022): "Online teaching and gender bias," *Economics of Education Review*, 89, 102280.

Ayllón, S., Á. Alsina, and J. Colomer (2019): "Teachers' involvement and students' self-efficacy: Keys to achievement in higher education," *PLoS ONE*, 14(5), e0216865.

Ayllón, S., L. Lefgren, R. Patterson, O. Stoddard, and N. Urdaneta (2024): "Equilibrium gender discrimination and disadvantage in student evaluations of teaching," Unpublished working paper.

Bagues, M., G. Vattuone, and N. Zinovyeva (2023): "Women in top academic positions: Is there a trickle-down effect?," Unpublished working paper.

Bayer, A., and C. Rouse (2016): "Diversity in the Economics profession: A new attack on an old problem," *Journal of Economic Perspectives*, 30(4), 221–242.

Beaman, L., R. Chattopadhyay, E. Duo, R. Pande, and P. Topalova (2009): "Powerful women: Does exposure reduce bias?," *The Quarterly Journal of Economics*, 124(4), 1497–1540.

Bertrand, M., D. Chugh, and S. Mullainathan (2005): "Implicit discrimination," *American Economic Review*, 95(2), 94–98.

Bertrand, M., and E. Duflo (2017): "Field experiments on discrimination," in *Handbook of Field Experiments*, ed. by A. V. Banerjee, and E. Duflo, vol. 1 of *Handbook of Economic Field Experiments*, chap. 8, pp. 309–393. North Holland, Elsevier.

Binderkrantz, A. S., M. Bisgaard, and B. Lassesen (2022): "Contradicting findings of gender bias in teaching evaluations: Evidence from two experiments in Denmark," *Assessment & Evaluation in Higher Education*, 47(8), 1345–1357.

Bohnet, I. (2016): *What Works: Gender equality by design*. Harvard University Press, Cambridge, MA.

Bohren, J. A., A. Imas, and M. Rosenberg (2019): "The dynamics of discrimination: Theory and evidence," *American Economic Review*, 109(10), 3395–3436.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016): "Stereotypes," *The Quarterly Journal of Economics*, 131(4), 1753–1794.

——— (2019): "Beliefs about gender," *American Economic Review*, 109(3), 739–773.

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

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

Boring, A., and A. Philippe (2021): "Reducing discrimination in the field: Evidence from an awareness raising intervention targeting gender biases in student evaluations of teaching," *Journal of Public Economics*, 193, 104323.

Card, D., F. Colella, and R. Lalive (2024): "Gender preferences in job vacancies and workplace gender diversity," *The Review of Economic Studies*, pp. n/a–n/a.

Card, D., S. DellaVigna, P. Funk, and N. Iriberri (2019): "Are referees and editors in Economics gender neutral?," *The Quarterly Journal of Economics*, 135(1), 269–327.

Carlana, M. (2019): "Implicit stereotypes: Evidence from teachers' gender bias," *The Quarterly Journal of Economics*, 134(3), 1163–1224.

Carnes, M., P. Devine, L. Baier, A. Byars-Winston, E. Fine, C. Ford, P. Forscher, C. Isaac, A. Kaatz, W. Magua, M. Palta, and J. Sheridan (2015): "The effect of an intervention to break the gender bias habit for faculty at one institution: A cluster randomized, controlled trial," *Academic Medicine*, 90(2), 221–230.

Ceci, S. J., S. Kahn, and W. M. Williams (2023): "Exploring gender bias in six key domains of academic science: An adversarial collaboration," *Psychological Science in the Public Interest*, 24(1), 15–73.

Chang, E. H., K. L. Milkman, D. M. Gromet, R. W. Rebele, C. Massey, A. L. Duckworth, and A. M. Grant (2019): "The mixed effects of online diversity training," *Proceedings of the National Academy of Sciences*, 116(16), 7778–7783.

Colby, G. (2023): "Data snapshot: Tenure and contingency in US higher education," Discussion paper, American Association of University Professors.

Delfino, A. (2024): "Breaking gender barriers: Experimental evidence on men in pink-collar jobs," *American Economic Review*, 114(6), 1816–1853.

Devine, P. G., P. S. Forscher, W. T. Cox, A. Kaatz, J. Sheridan, and M. Carnes (2017): "A gender bias habit-breaking intervention led to increased hiring of female faculty in STEMM departments," *Journal of Experimental Social Psychology*, 73, 211–215.

Dobbin, F., and A. Kalev (2016): "Why diversity programs fail," *Harvard Business Review*, 94(7), n/a–n/a.

——— (2018): "Why doesn't diversity training work? The challenge for industry and academia," *Anthropology Now*, 10, 48–55.

Eberhardt, M., G. Facchini, and V. Rueda (2023): "Gender differences in reference letters: Evidence from the economics job market," *The Economic Journal*, 133(655), 2676–2708.

European Commission and Directorate-General for Research and Innovation (2021): *She Figures 2021: Gender in research and innovation: Statistics and indicators*. Luxembourg: Publications Office of the European Union.

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

Fisk, S., K. Stolee, and L. Battestilli (2020): "A lightweight intervention to decrease gender bias in student evaluations of teaching," *Research in Equity and Sustained Participation in Engineering, Computing, and Technology*, 1, 1–4.

Genetin, B., J. Chen, V. Kogan, and A. Kalish (2022): "Mitigating implicit bias in student evaluations: A randomized intervention," *Applied Economic Perspectives and Policy*, 44(1), 110–128.

Glover, D., A. Pallais, and W. Pariente (2017): "Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores," *The Quarterly Journal of Economics*, 132(3), 1219–1260.

Goldin, C., and C. Rouse (2000): "Orchestrating impartiality: The impact of 'blind' auditions on female musicians," *The American Economic Review*, 90(4), 715–741.

Hengel, E. (2022): "Publishing while female: Are women held to higher standards? Evidence from peer review," *The Economic Journal*, 132(648), 2951–2991.

Hoorens, V., G. Dekkers, and E. Deschrijver (2021): "Gender bias in student evaluations of teaching: Students' self-affirmation reduces the bias by lowering evaluations of male professors," *Sex Roles*, 84, 34–48.

Kalev, A., F. Dobbin, and E. Kelly (2006): "Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies," *American Sociological Review*, 71(4), 589–617.

Keng, S.-H. (2020): "Gender bias and statistical discrimination against female instructors in student evaluations of teaching," *Labour Economics*, 66, 101889.

Kunze, A., and A. Miller (2017): "Women helping women? Evidence from private sector data on workplace hierarchies," *The Review of Economics and Statistics*, 99(5), 769–775.

Lai, C. K., K. M. Hoffman, and B. A. Nosek (2013): "Reducing implicit prejudice," *Social and Personality Psychology Compass*, 7(5), 315–330.

Lai, C. K., M. Marini, S. A. Lehr, C. Cerruti, J.-E. L. Shin, J. A. Joy-Gaba, A. K. Ho, B. A. Teachman, S. P. Wojcik, S. P. Koleva, R. S. Frazier, L. Heiphetz, E. E. Chen, R. N. Turner, J. Haidt, S. Kesebir, C. B. Hawkins, H. S. Schaefer, S. Rubichi, G. Sartori, C. M. Dial, N. Sriram, M. R. Banaji, and B. A. Nosek (2014): "Reducing implicit racial preferences: I. A comparative investigation of 17 interventions," *Journal of Experimental Psychology*, 143(4), 1765–1785.

Lai, C. K., A. Skinner, E. Cooley, S. Murrar, M. Brauer, T. Devos, J. Calanchini, Y. Xiao, C. Pedram, C. Marshburn, S. Simon, J. Blanchar, J. Joy-Gaba, J. Conway, L. Redford, R. Klein, G. Roussos, F. Schellhaas, M. Burns, X. Hu, M. McLean, J. Axt, S. Asgari, K. Schmidt, R. Rubinstein, M. Marini, S. Rubichi, J. Shin, and B. Nosek (2016): "Reducing implicit racial preferences: II. Intervention effectiveness across time," *Journal of Experimental Psychology*, 145(8), 1001.

Legault, L., J. N. Gutsell, and M. Inzlicht (2011): "Ironic effects of antiprejudice messages: How motivational interventions can reduce (but also increase) prejudice," *Psychological Science*, 22(12), 1472–1477.

Lincoln, A. E., S. H. Pincus, J. Koster, and P. S. Leboy (2012): "The Matilda Effect in science: Awards and prizes in the US, 1990s and 2000s," *Social Studies of Science*, 42(2), 307–320.

MacNell, L., A. Driscoll, and A. Hunt (2015): "What's in a name: Exposing gender bias in student ratings of teaching," *Innovative Higher Education*, 40, 291–303.

McElroy, M. (2016): "Report: Committee on the Status of Women in the Economics Profession (CSWEP)," *American Economic Review*, 106(5), 750–773.

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

Miron, A. C., and J. W. Brehm (2006): "Reactance theory – 40 years later," *Zeitschrift für Sozialpsychologie / Journal of Psychology*, 37, 9–18.

Mitchell, K. M. W., and J. Martin (2018): "Gender bias in student evaluations," *Political Science & Politics*, 51(3), 648–652.

Moss-Racusin, C. A., E. S. Pietri, E. P. Hennes, J. F. Dovidio, B. L. Brescoll, G. Roussos, and J. Handelsman (2018): "Reducing STEM gender bias with VIDS (video interventions for diversity in STEM)," *Journal of Experimental Psychology: Applied*, 24(2), 236–260.

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

Mühlberger, C., and E. Jonas (2019): "Reactance theory," in *Social Psychology in Action*, chap. 6, pp. 79–94. Cham: Springer International Publishing.

Neckermann, S., U. Turmunkh, D. van Dolder, and T. V. Wang (2022): "Nudging student participation in online evaluations of teaching: Evidence from a field experiment," *European Economic Review*, 141, 104001.

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.

Peterson, D., L. Biederman, D. Andersen, T. Ditonto, and K. Roe (2019): "Mitigating gender bias in student evaluations of teaching," *PLoS ONE*, 14(5), e0216241.

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

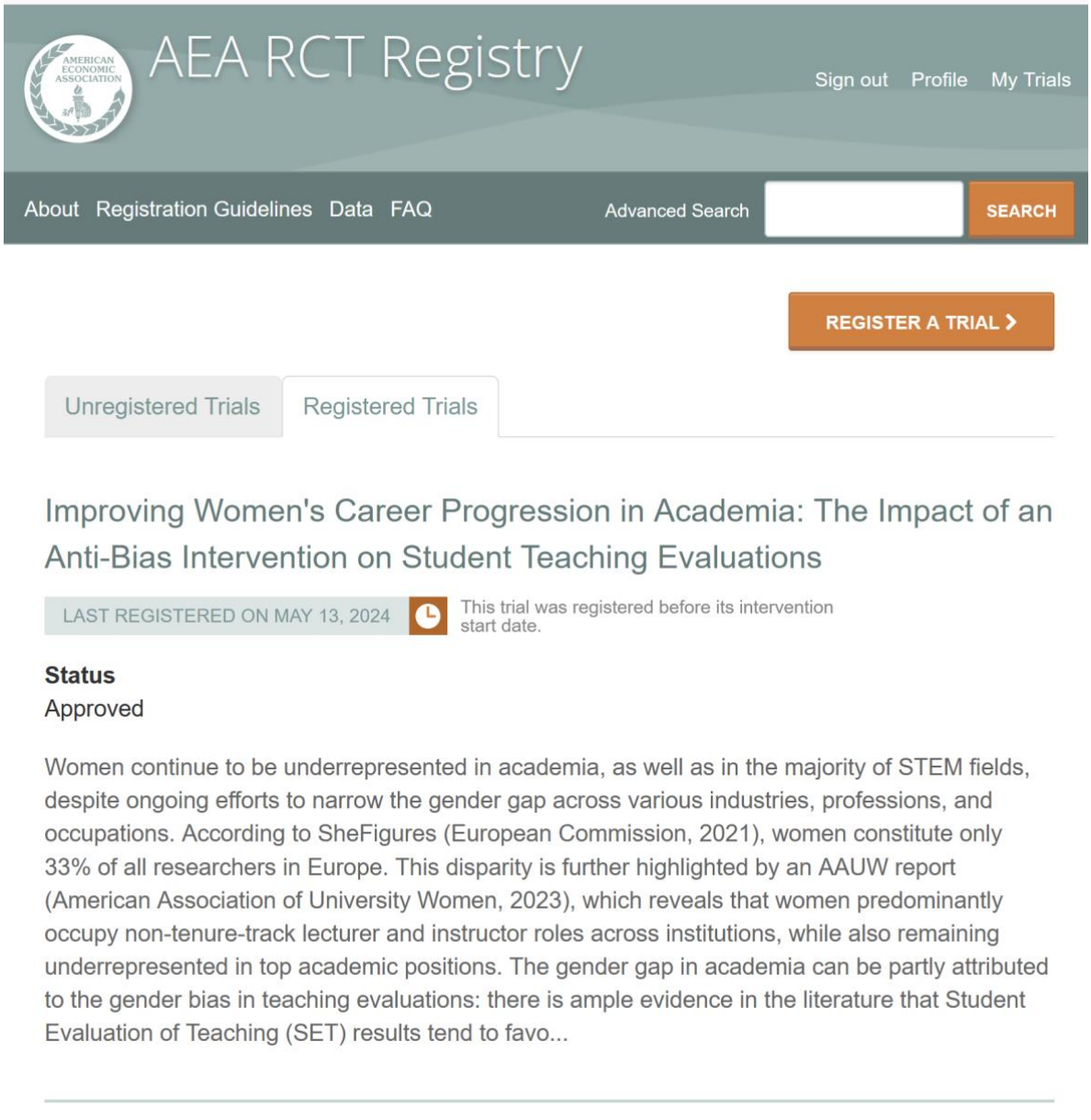
Rooth, D. O. (2010): "Automatic associations and discrimination in hiring: Real world evidence," *Labour Economics*, 17(3), 523–534.

Silver, J. K., C. Slocum, A. M. Bank, S. Bhatnagar, C. A. Blauwet, J. A. Poorman, A. C. Villablanca, and S. Parangi (2017): "Where are the women? The underrepresentation of women physicians among recognition award recipients from medical specialty societies," *Physical Medicine & Rehabilitation Journal*, 9(8), 804–815.

Sinclair, L., and Z. Kunda (2000): "Motivated stereotyping of women: She's fine if she praised me but incompetent if she criticized me," *Personality and Social Psychology Bulletin*, 26(11), 1329–1342.

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

A American Economic Association RCT Registry approval and field experiment approval by the Ethics Committee at the University of Girona



DICTAMEN DEL COMITÈ D'ÈTICA I BIOSEGURETAT DE LA RECERCA DE LA UNIVERSITAT DE GIRONA

Nom projecte: Sesgo de genero en las encuestas docentes

Codi projecte: CEBRU0002-24

Investigadora: Sara Ayllon Gatnau

Helena Montiel Boadas, secretària del Comitè d'Ètica i Bioseguretat de la Recerca de la Universitat de Girona,

FAIG CONSTAR :

Que en la sessió ordinària número 3/2024 que va tenir lloc el dia 18 de març de 2024, el Comitè d'Ètica i Bioseguretat de la Recerca de la Universitat de Girona va avaluar el protocol del projecte "Sesgo de genero en las encuestas docentes" i va considerar per unanimitat que compleix els requeriments ètics i de bioseguretat exigibles.

Que és responsabilitat dels investigadors que la recerca es realitzi tal i com descriu la documentació presentada. Qualsevol canvi significatiu ha de ser comunicat al Comitè, la qual cosa requerirà una nova valoració.

Per la qual cosa, s'emet aquest dictamen favorable.

Helena Montiel Boadas

Secretària del Comitè d'Ètica i Bioseguretat de la Recerca de la Universitat de Girona

Table 3: Robustness checks

	All students		Male students		Female students	
	Male lecturer	Female lecturer	Male lecturer	Female lecturer	Male lecturer	Female lecturer
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: FE by professor						
Treatment 1	-0.071***	-0.027	-0.044	-0.018	-0.075**	-0.018
	(0.02)	(0.02)	(0.04)	(0.04)	(0.03)	(0.03)
Treatment 2	-0.011	-0.031	0.014	-0.119**	-0.017	0.005
	(0.02)	(0.02)	(0.04)	(0.05)	(0.03)	(0.03)
Observations	13445	12682	5821	3739	7494	8826
R2	0.316	0.314	0.353	0.346	0.350	0.341
Panel B: Clustered standard errors by teacher						
Treatment 1	-0.069***	-0.027	-0.068*	-0.037	-0.064**	-0.009
	(0.02)	(0.02)	(0.04)	(0.04)	(0.03)	(0.03)
Treatment 2	-0.020	-0.031	-0.020	-0.128***	-0.020	0.011
	(0.02)	(0.02)	(0.04)	(0.05)	(0.03)	(0.03)
Observations	13281	12492	5641	3582	7324	8636
R2	0.363	0.361	0.407	0.394	0.403	0.396
Panel C: Responses 1, 2 vs. 3, 4, 5						
Treatment 1	-0.008	-0.008	-0.009	-0.010	-0.005	-0.005
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Treatment 2	-0.004	-0.006	-0.010	-0.035**	0.000	0.009
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Observations	13281	12492	5641	3582	7324	8636
R2	0.262	0.271	0.314	0.318	0.296	0.304
Panel D: Responses 1, 2, 3 , 4 vs. 5						
Treatment 1	-0.038***	-0.009	-0.038**	-0.025	-0.035**	0.002
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Treatment 2	-0.010	-0.011	0.006	-0.026	-0.019	-0.009
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Observations	13281	12492	5641	3582	7324	8636
R2	0.295	0.274	0.335	0.324	0.343	0.304
Panel E: Minimum number of evaluations						
Treatment 1	-0.071**	-0.020	-0.105**	-0.016	-0.045	-0.005
	(0.03)	(0.03)	(0.05)	(0.06)	(0.04)	(0.03)
Treatment 2	-0.010	-0.046	-0.012	-0.162**	-0.015	-0.003
	(0.03)	(0.03)	(0.05)	(0.06)	(0.04)	(0.03)
Observations	7421	7142	3127	2026	4284	5103
R2	0.330	0.321	0.359	0.333	0.363	0.354

Note: Each column in each panel is the result of a different regression, where the dependent variable is the teaching evaluation score given to lecturer j by student i relative to course s. All regressions include student characteristics (age, age-squared and final grade obtained) and professor- and course-level fixed effects — except in Panel A, which does not include course-level fixed effects. Columns (1) and (2) also control for students' gender. The number of observations in columns (3) and (5) may not add up to those of column (1) because the number of singleton observations in each regression may differ. The same is true in the case of columns (4) and (6), in relation to column (2). Robust standard errors in parentheses, except in Panel (B). *** significant at 1%, ** at 5% and * at 10%.

Source: Authors' computation, using data from the University of Girona.

Table 4: Heterogeneous results by field of knowledge

	All students		Male students		Female students	
	Male lecturer (1)	Female lecturer (2)	Male lecturer (3)	Female lecturer (4)	Male lecturer (5)	Female lecturer (6)
Panel A: Arts and Humanities						
Treatment 1	-0.000	-0.180	-0.362	-0.510**	0.074	-0.104
	(0.13)	(0.13)	(0.22)	(0.21)	(0.22)	(0.20)
Treatment 2	0.009	-0.051	-0.050	-0.173	-0.181	-0.166
	(0.12)	(0.10)	(0.19)	(0.15)	(0.22)	(0.20)
Observations	601	624	266	260	287	343
R2	0.437	0.430	0.455	0.541	0.505	0.402
Panel B: Sciences						
Treatment 1	-0.040	0.030	-0.089	-0.004	-0.011	0.061
	(0.04)	(0.04)	(0.08)	(0.08)	(0.05)	(0.05)
Treatment 2	0.004	0.049	0.041	-0.041	-0.023	0.091*
	(0.05)	(0.04)	(0.09)	(0.09)	(0.06)	(0.05)
Observations	2547	2877	866	927	1638	1916
R2	0.336	0.265	0.377	0.276	0.370	0.313
Panel C: Life Sciences						
Treatment 1	-0.097*	0.031	0.404***	0.001	-0.164***	0.038
	(0.05)	(0.04)	(0.15)	(0.10)	(0.06)	(0.05)
Treatment 2	-0.157***	-0.056	-0.220	-0.574***	-0.143**	0.002
	(0.06)	(0.04)	(0.16)	(0.11)	(0.06)	(0.04)
Observations	1853	3500	333	527	1486	2913
R2	0.354	0.308	0.511	0.499	0.368	0.317
Panel D: Social Sciences						
Treatment 1	-0.140***	-0.109**	-0.281***	-0.115	-0.065	-0.107**
	(0.04)	(0.05)	(0.08)	(0.11)	(0.06)	(0.05)
Treatment 2	-0.098**	-0.072	-0.251***	-0.231*	-0.019	-0.011
	(0.04)	(0.05)	(0.09)	(0.12)	(0.05)	(0.05)
Observations	3879	3701	1168	827	2598	2743
R2	0.371	0.420	0.478	0.460	0.399	0.455
Panel E: Architecture and Engineering						
Treatment 1	-0.012	-0.032	-0.016	0.026	-0.014	-0.049
	(0.04)	(0.06)	(0.05)	(0.09)	(0.08)	(0.09)
Treatment 2	0.089**	-0.041	0.059	0.017	0.186**	-0.109
	(0.04)	(0.07)	(0.05)	(0.10)	(0.08)	(0.10)
Observations	4397	1784	3001	1032	1300	716
R2	0.350	0.381	0.372	0.358	0.441	0.518

Note: Each column in each panel is the result of a different regression, where the dependent variable is the teaching evaluation score given to lecturer j by student i relative to course s. All regressions include student characteristics (age, age-squared and final grade obtained) and professor- and course-level fixed effects. Columns (1) and (2) also control for students' gender. The number of observations in columns (3) and (5) may not add up to those of column (1) because the number of singleton observations in each regression may differ. The same is true in the case of columns (4) and (6), in comparison with column (2). Robust standard errors in parentheses. *** significant at 1%, ** at 5% and * at 10%.

Source: Authors' computation, using data from the University of Girona.

Table 5: Heterogeneous results by faculty — continues on next page

	All students		Male students		Female students	
	Male lecturer	Female lecturer	Male lecturer	Female lecturer	Male lecturer	Female lecturer
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Law Faculty						
Treatment 1	-0.125	-0.053	0.011	0.101	-0.242	-0.082
	(0.10)	(0.12)	(0.14)	(0.24)	(0.15)	(0.15)
Treatment 2	-0.134	-0.207	-0.141	-0.209	-0.123	-0.062
	(0.10)	(0.13)	(0.15)	(0.27)	(0.14)	(0.15)
Observations	813	480	298	159	485	303
R2	0.394	0.445	0.534	0.489	0.419	0.522
Panel B: Economics Faculty						
Treatment 1	-0.286***	0.026	-0.430***	-0.262	-0.016	0.398**
	(0.09)	(0.12)	(0.11)	(0.17)	(0.13)	(0.18)
Treatment 2	-0.191**	-0.105	-0.282**	-0.249	-0.009	0.048
	(0.09)	(0.12)	(0.13)	(0.18)	(0.13)	(0.18)
Observations	1106	629	559	310	518	302
R2	0.335	0.314	0.426	0.375	0.404	0.318
Panel C: Sciences Faculty						
Treatment 1	-0.040	0.030	-0.089	-0.004	-0.011	0.061
	(0.04)	(0.04)	(0.08)	(0.08)	(0.05)	(0.05)
Treatment 2	0.004	0.049	0.041	-0.041	-0.023	0.091*
	(0.05)	(0.04)	(0.09)	(0.09)	(0.06)	(0.05)
Observations	2547	2877	866	927	1638	1916
R2	0.336	0.265	0.377	0.276	0.370	0.313
Panel D: Technology Polythecnic						
Treatment 1	-0.012	-0.032	-0.016	0.026	-0.014	-0.049
	(0.04)	(0.06)	(0.05)	(0.09)	(0.08)	(0.09)
Treatment 2	0.089**	-0.041	0.059	0.017	0.186**	-0.109
	(0.04)	(0.07)	(0.05)	(0.10)	(0.08)	(0.10)
Observations	4397	1784	3001	1032	1300	716
R2	0.350	0.381	0.372	0.358	0.441	0.518
Panel E: Education and Psychology Faculty						
Treatment 1	-0.187***	-0.180***	0.033	0.066	-0.235***	-0.191***
	(0.07)	(0.05)	(0.20)	(0.19)	(0.08)	(0.06)
Treatment 2	-0.154**	-0.100**	-0.259	-0.384**	-0.160**	-0.049
	(0.07)	(0.05)	(0.22)	(0.17)	(0.07)	(0.05)
Observations	1559	2644	216	323	1301	2242
R2	0.376	0.470	0.526	0.581	0.382	0.480

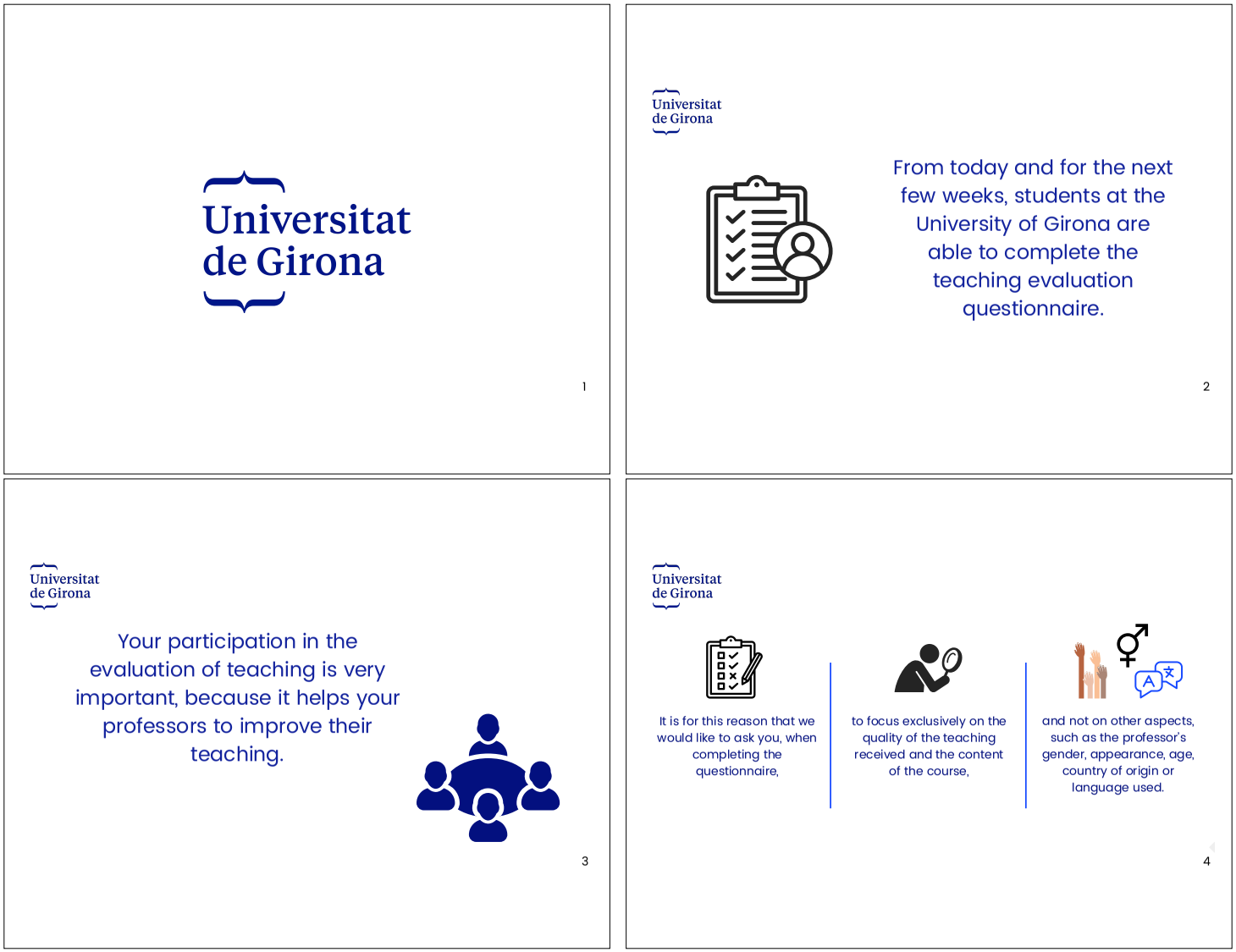
Table 5: Heterogeneous results by faculty — continued from previous page

	All students		Male students		Female students	
	Male lecturer	Female lecturer	Male lecturer	Female lecturer	Male lecturer	Female lecturer
	(1)	(2)	(3)	(4)	(5)	(6)
Panel F: Nursing Faculty						
Treatment 1	0.019	0.060	0.434*	-0.206	-0.050	0.093*
	(0.09)	(0.05)	(0.25)	(0.14)	(0.10)	(0.06)
Treatment 2	-0.127	-0.075	-0.645**	-0.768***	-0.114	-0.023
	(0.09)	(0.05)	(0.28)	(0.16)	(0.10)	(0.06)
Observations	785	2035	100	230	677	1785
R2	0.249	0.144	0.492	0.347	0.244	0.145
Panel G: Medicine Faculty						
Treatment 1	-0.179**	-0.075	0.477**	0.292	-0.235**	-0.152
	(0.08)	(0.11)	(0.24)	(0.18)	(0.10)	(0.15)
Treatment 2	-0.167*	0.013	0.014	-0.103	-0.139	0.031
	(0.09)	(0.10)	(0.20)	(0.20)	(0.11)	(0.13)
Observations	721	636	196	184	504	417
R2	0.378	0.333	0.514	0.549	0.444	0.393
Panel H: Tourism Faculty						
Treatment 1	0.145	0.059	0.006	0.216	0.172	0.027
	(0.10)	(0.10)	(0.28)	(0.26)	(0.12)	(0.12)
Treatment 2	0.128	0.155	0.096	-0.061	0.180*	0.310**
	(0.09)	(0.11)	(0.30)	(0.33)	(0.10)	(0.12)
Observations	745	766	129	142	591	595
R2	0.432	0.399	0.578	0.521	0.454	0.464
Panel I: Arts and Humanities Faculty						
Treatment 1	-0.000	-0.180	-0.362	-0.510**	0.074	-0.104
	(0.13)	(0.13)	(0.22)	(0.21)	(0.22)	(0.20)
Treatment 2	0.009	-0.051	-0.050	-0.173	-0.181	-0.166
	(0.12)	(0.10)	(0.19)	(0.15)	(0.22)	(0.20)
Observations	601	624	266	260	287	343
R2	0.437	0.430	0.455	0.541	0.505	0.402

Note: Each column in each panel is the result of a different regression, where the dependent variable is the teaching evaluation score given to lecturer j by student i relative to course s. All regressions include student characteristics (age, age-squared and final grade obtained) and professor- and course-level fixed effects. Columns (1) and (2) also control for students' gender. The number of observations in columns (3) and (5) may not add up to those of column (1) because the number of singleton observations in each regression may differ. The same is true in the case of columns (4) and (6), in comparison with column (2). Robust standard errors in parentheses. *** significant at 1%, ** at 5% and * at 10%.
Source: Authors' computation, using data from the University of Girona.

B Video treatments

Treatment 1


Universitat de Girona
1
Universitat de Girona
From today and for the next few weeks, students at the University of Girona are able to complete the teaching evaluation questionnaire.
2
Universitat de Girona
Your participation in the evaluation of teaching is very important, because it helps your professors to improve their teaching.
3
Universitat de Girona
It is for this reason that we would like to ask you, when completing the questionnaire,
to focus exclusively on the quality of the teaching received and the content of the course,
and not on other aspects, such as the professor's gender, appearance, age, country of origin or language used.
4

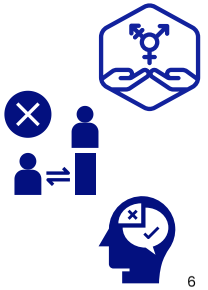
Universitat de Girona

In this regard, it is important to avoid implicit bias which shows prejudice, normally negative, towards a given group -- bias of which we are not conscious and which often arises involuntarily.

5

Universitat de Girona

We ask you, therefore, to avoid prejudice or discriminatory behaviour when filling in the questionnaire.



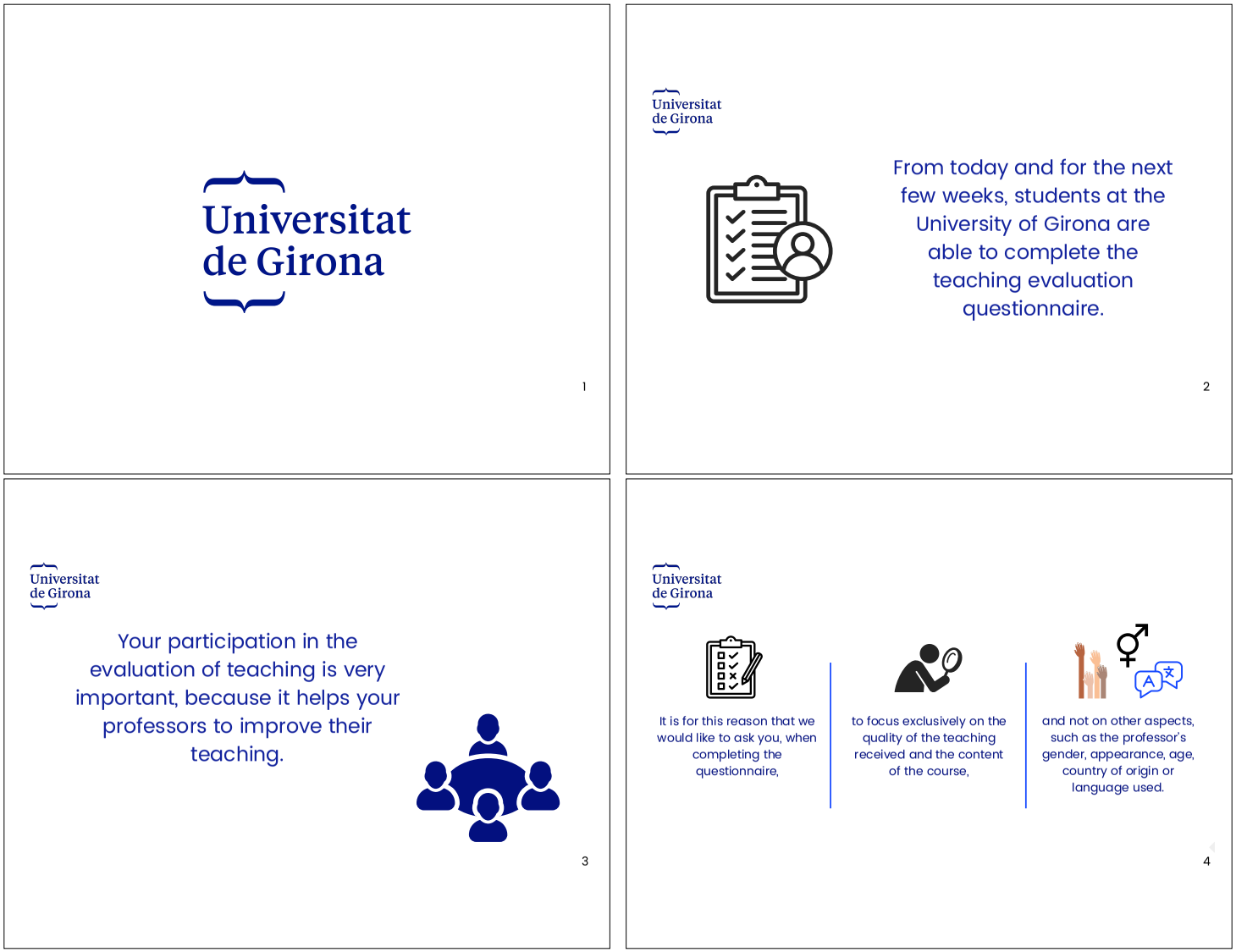
6

Universitat de Girona

Thank you!

7

Treatment 2

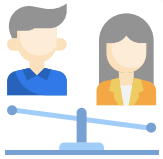

Universitat de Girona
1
Universitat de Girona
From today and for the next few weeks, students at the University of Girona are able to complete the teaching evaluation questionnaire.
2
Universitat de Girona
Your participation in the evaluation of teaching is very important, because it helps your professors to improve their teaching.
3
Universitat de Girona
It is for this reason that we would like to ask you, when completing the questionnaire,
to focus exclusively on the quality of the teaching received and the content of the course,
and not on other aspects, such as the professor's gender, appearance, age, country of origin or language used.
4

Universitat de Girona

The academic literature has shown that, on average, female professors receive lower scores for their teaching compared to male professors, even when the quality of their teaching is the same.

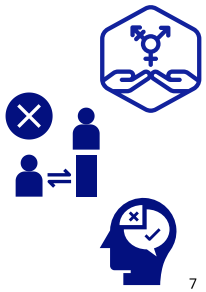
Universitat de Girona

Additionally, it has been documented that such biased behaviour often arises from male students (and not so much from female students).



Universitat de Girona

We ask you, therefore, to avoid prejudice or discriminatory behaviour when filling in the questionnaire.



Universitat de Girona

Thank you!

C Additional Tables and Figures

Table C.1: Mean differences across treatment arms

	Control	T1	T2	C vs. T1 p-value	C vs. T2 p-value	T1 vs. T2 p-value
Female student	0.573	0.571	0.571	0.840	0.856	0.983
	(0.494)	(0.494)	(0.494)			
Law Faculty	0.096	0.094	0.096	0.693	0.918	0.770
	(0.295)	(0.292)	(0.294)			
Economics Faculty	0.088	0.094	0.098	0.290	0.083	0.501
	(0.283)	(0.292)	(0.298)			
Sciences Faculty	0.102	0.095	0.089	0.298	0.036	0.292
	(0.302)	(0.294)	(0.285)			
Technology Polytechnic	0.174	0.177	0.177	0.742	0.707	0.962
	(0.379)	(0.382)	(0.382)			
Education and Psychology Faculty	0.157	0.151	0.152	0.456	0.490	0.956
	(0.364)	(0.358)	(0.359)			
Nursing Faculty	0.048	0.044	0.051	0.326	0.523	0.105
	(0.215)	(0.206)	(0.221)			
Medicine Faculty	0.040	0.035	0.033	0.157	0.046	0.562
	(0.197)	(0.184)	(0.178)			
Tourism Faculty	0.052	0.053	0.050	0.785	0.562	0.394
	(0.223)	(0.225)	(0.217)			
Arts and Humanities Faculty	0.053	0.056	0.058	0.557	0.234	0.547
	(0.224)	(0.229)	(0.235)			
Professional schools	0.184	0.196	0.191	0.162	0.384	0.596
	(0.388)	(0.397)	(0.393)			
Observations	4,731	4,713	4,720			

Note: The table shows mean differences across treatment arms prior to the intervention. The unit of analysis is the student level. Standard errors in parentheses.
Source: Authors' computation, using data from the University of Girona.

Table C.2: Heterogeneous results by lecturer's age

	All students		Male students		Female students	
	Male lecturer	Female lecturer	Male lecturer	Female lecturer	Male lecturer	Female lecturer
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Less than 35 years of age						
Treatment 1	-0.074	-0.147**	0.017	-0.129	-0.184*	-0.106
	(0.07)	(0.07)	(0.10)	(0.14)	(0.09)	(0.08)
Treatment 2	-0.076	-0.038	-0.129	-0.119	0.016	-0.006
	(0.07)	(0.06)	(0.11)	(0.15)	(0.09)	(0.08)
Observations	1467	1399	692	332	748	1036
R2	0.382	0.313	0.416	0.479	0.444	0.312
Panel B: Between 35 and 49						
Treatment 1	-0.025	-0.010	-0.019	-0.022	-0.041	-0.007
	(0.04)	(0.03)	(0.06)	(0.07)	(0.05)	(0.04)
Treatment 2	-0.009	-0.017	0.020	-0.163**	-0.051	0.030
	(0.04)	(0.03)	(0.07)	(0.07)	(0.05)	(0.04)
Observations	4992	5622	1914	1486	2945	4010
R2	0.322	0.323	0.363	0.375	0.367	0.345
Panel C: More than 50 years of age						
Treatment 1	-0.116***	-0.027	-0.135***	-0.057	-0.082*	0.010
	(0.03)	(0.04)	(0.05)	(0.07)	(0.05)	(0.05)
Treatment 2	-0.031	-0.079**	-0.035	-0.158**	-0.010	-0.050
	(0.03)	(0.04)	(0.05)	(0.07)	(0.05)	(0.05)
Observations	6116	4804	2728	1575	3252	3128
R2	0.391	0.369	0.439	0.360	0.425	0.428

Note: Each column in each panel is the result of a different regression, where the dependent variable is the teaching evaluation score given to lecturer j by student i relative to course s. All regressions include student characteristics (age, age-squared and final grade obtained) and professor- and course-level fixed effects. Columns (1) and (2) also control for students' gender. The number of observations in columns (3) and (5) may not add up to those of column (1) because the number of singleton observations in each regression may differ. The same is true in the case of columns (4) and (6), in comparison with column (2). Robust standard errors in parentheses. *** significant at 1%, ** at 5% and * at 10%.
Source: Authors' computation, using data from the University of Girona.

Table C.3: Heterogeneous results by professor's tenure

	All students		Male students		Female students	
	Male lecturer	Female lecturer	Male lecturer	Female lecturer	Male lecturer	Female lecturer
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Full professor or emeritus						
Treatment 1	-0.043	0.018	-0.032	-0.116	-0.007	-0.011
	(0.08)	(0.11)	(0.12)	(0.25)	(0.11)	(0.13)
Treatment 2	0.079	-0.009	0.126	0.076	0.042	-0.080
	(0.08)	(0.11)	(0.13)	(0.21)	(0.11)	(0.15)
Observations	1269	613	571	182	667	421
R2	0.323	0.311	0.401	0.376	0.322	0.338
Panel B: Associate professor						
Treatment 1	-0.138***	-0.049	-0.201***	-0.048	-0.088	-0.036
	(0.04)	(0.04)	(0.06)	(0.08)	(0.05)	(0.05)
Treatment 2	0.003	-0.054	-0.058	-0.146*	0.057	-0.011
	(0.04)	(0.04)	(0.06)	(0.08)	(0.06)	(0.05)
Observations	4289	3791	1950	1332	2219	2397
R2	0.408	0.406	0.426	0.387	0.450	0.467
Panel C: Assistant or visiting professor						
Treatment 1	-0.119	-0.058	-0.129	0.088	-0.125	-0.106
	(0.09)	(0.07)	(0.15)	(0.19)	(0.14)	(0.08)
Treatment 2	-0.010	-0.055	0.011	-0.146	-0.048	-0.048
	(0.09)	(0.07)	(0.16)	(0.20)	(0.13)	(0.08)
Observations	914	1221	458	211	433	986
R2	0.338	0.381	0.309	0.365	0.445	0.407
Panel D: Adjunct or pre-PhD faculty						
Treatment 1	-0.024	-0.012	0.034	-0.035	-0.054	0.027
	(0.03)	(0.03)	(0.05)	(0.07)	(0.04)	(0.04)
Treatment 2	-0.056*	-0.014	-0.037	-0.133*	-0.067*	0.040
	(0.03)	(0.03)	(0.06)	(0.07)	(0.04)	(0.04)
Observations	6809	6867	2662	1857	4005	4832
R2	0.343	0.328	0.415	0.403	0.379	0.349

Note: Each column in each panel is the result of a different regression, where the dependent variable is the teaching evaluation score given to lecturer j by student i relative to course s. All regressions include student characteristics (age, age-squared and final grade obtained) and professor- and course-level fixed effects. Columns (1) and (2) also control for students' gender. The number of observations in columns (3) and (5) may not add up to those of column (1) because the number of singleton observations in each regression may differ. The same is true in the case of columns (4) and (6), in comparison with column (2). Robust standard errors in parentheses. *** significant at 1%, ** at 5% and * at 10%.
Source: Authors' computation, using data from the University of Girona.

Table C.4: Heterogeneous results by student's final grade obtained

	All students		Male students		Female students	
	Male lecturer	Female lecturer	Male lecturer	Female lecturer	Male lecturer	Female lecturer
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Failed (< 5)						
Treatment 1	-0.138	-0.044	-0.141	-0.184	-0.026	0.057
	(0.11)	(0.13)	(0.16)	(0.25)	(0.20)	(0.19)
Treatment 2	0.021	0.139	0.123	-0.234	-0.179	0.259
	(0.11)	(0.12)	(0.14)	(0.22)	(0.19)	(0.17)
Observations	883	618	445	222	317	332
R2	0.402	0.440	0.488	0.427	0.425	0.501
Panel B: Passed (5 to 6.9)						
Treatment 1	-0.111***	-0.012	-0.126*	0.040	-0.086	-0.019
	(0.04)	(0.05)	(0.06)	(0.09)	(0.06)	(0.07)
Treatment 2	0.001	-0.104**	-0.069	-0.155	0.060	-0.009
	(0.04)	(0.05)	(0.07)	(0.10)	(0.06)	(0.07)
Observations	4647	3092	2144	1086	2267	1830
R2	0.385	0.395	0.420	0.411	0.461	0.451
Panel C: Good (7 to 8.9)						
Treatment 1	-0.031	-0.046	-0.013	-0.166**	-0.037	-0.014
	(0.03)	(0.03)	(0.06)	(0.07)	(0.05)	(0.04)
Treatment 2	-0.029	-0.008	-0.038	-0.233***	-0.024	0.059
	(0.03)	(0.03)	(0.07)	(0.08)	(0.04)	(0.04)
Observations	5637	6510	1938	1519	3340	4681
R2	0.404	0.379	0.495	0.431	0.424	0.414
Panel D: Excellent (>= 9)						
Treatment 1	-0.121*	0.061	-0.008	-0.224	-0.286***	0.088
	(0.07)	(0.06)	(0.12)	(0.18)	(0.10)	(0.08)
Treatment 2	-0.003	0.055	0.219	-0.213	-0.289***	0.050
	(0.07)	(0.06)	(0.14)	(0.16)	(0.09)	(0.07)
Observations	1311	1646	417	234	690	1211
R2	0.515	0.560	0.581	0.490	0.578	0.596

Note: Each column in each panel is the result of a different regression, where the dependent variable is the teaching evaluation score given to lecturer j by student i relative to course s. All regressions include student characteristics (age, age-squared and final grade obtained) and professor- and course-level fixed effects. Columns (1) and (2) also control for students' gender. The number of observations in columns (3) and (5) may not add up to those of column (1) because the number of singleton observations in each regression may differ. The same is true in the case of columns (4) and (6), in comparison with column (2). Robust standard errors in parentheses. *** significant at 1%, ** at 5% and * at 10%.
Source: Authors' computation, using data from the University of Girona.

Table C.5: Heterogeneous results by student's seniority

	All students		Male students		Female students	
	Male lecturer	Female lecturer	Male lecturer	Female lecturer	Male lecturer	Female lecturer
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 1st. year students						
Treatment 1	-0.106***	-0.083**	-0.060	-0.005	-0.127***	-0.093*
	(0.04)	(0.04)	(0.06)	(0.08)	(0.05)	(0.05)
Treatment 2	0.037	-0.041	0.053	-0.023	0.052	-0.030
	(0.04)	(0.04)	(0.06)	(0.09)	(0.05)	(0.05)
Observations	4935	3908	2094	1199	2780	2657
R2	0.341	0.334	0.378	0.351	0.389	0.362
Panel B: 2nd. year students						
Treatment 1	-0.033	0.039	-0.086	-0.050	0.006	0.071
	(0.05)	(0.04)	(0.07)	(0.09)	(0.06)	(0.05)
Treatment 2	-0.005	-0.028	0.030	-0.279***	-0.050	0.035
	(0.05)	(0.04)	(0.08)	(0.10)	(0.06)	(0.05)
Observations	3556	3791	1515	916	2004	2813
R2	0.325	0.301	0.396	0.383	0.346	0.324
Panel C: 3rd. year students						
Treatment 1	-0.088*	-0.033	-0.116	0.004	-0.061	-0.022
	(0.05)	(0.05)	(0.08)	(0.09)	(0.07)	(0.06)
Treatment 2	-0.068	-0.022	-0.095	0.001	-0.048	0.012
	(0.05)	(0.05)	(0.08)	(0.11)	(0.07)	(0.06)
Observations	2575	2424	1200	790	1312	1604
R2	0.385	0.422	0.431	0.423	0.421	0.480
Panel D: 4th. and 5th year students						
Treatment 1	-0.042	-0.234**	-0.126	-0.310*	-0.036	-0.226*
	(0.09)	(0.09)	(0.16)	(0.16)	(0.11)	(0.13)
Treatment 2	-0.131	-0.230**	-0.201	-0.440***	-0.159	-0.266*
	(0.09)	(0.10)	(0.17)	(0.15)	(0.11)	(0.14)
Observations	876	752	340	266	486	454
R2	0.418	0.324	0.470	0.389	0.496	0.347

Note: Each column in each panel is the result of a different regression, where the dependent variable is the teaching evaluation score given to lecturer j by student i relative to course s. All regressions include student characteristics (age, age-squared and final grade obtained) and professor- and course-level fixed effects. Columns (1) and (2) also control for students' gender. The number of observations in columns (3) and (5) may not add up to those of column (1) because the number of singleton observations in each regression may differ. The same is true in the case of columns (4) and (6), in comparison with column (2). Robust standard errors in parentheses. *** significant at 1%, ** at 5% and * at 10%.
Source: Authors' computation, using data from the University of Girona.

Table C.6: Sample composition

	Student female	Student age	Student grade	Arts and Humanities	Sciences	Life Sciences	Social Sciences	Engineering
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Male lecturers								
Treatment 1	-0.027***	-0.262***	0.052*	-0.002**	0.000	-0.000	0.005***	-0.003**
	(0.01)	(0.09)	(0.03)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Treatment 2	0.021**	0.065	0.040	-0.000	0.000	-0.000	0.005***	-0.004***
	(0.01)	(0.09)	(0.03)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Observations	13281	13281	13281	13281	13281	13281	13281	13281
R2	0.296	0.401	0.418	0.985	1.000	1.000	0.986	0.990
Panel B: Female lecturers								
Treatment 1	-0.024**	-0.429***	-0.065**	0.000*	-0.000	0.000	0.001	-0.001
	(0.01)	(0.11)	(0.03)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Treatment 2	-0.023**	0.107	-0.018	-0.000	-0.000	0.000	0.002**	-0.002*
	(0.01)	(0.11)	(0.03)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Observations	12492	12492	12492	12492	12492	12492	12492	12492
R2	0.243	0.292	0.472	0.994	1.000	1.000	0.992	0.988

Note: Each column in each panel is the result of a different regression (following Equation (1)), where the dependent variable is specified in the column header. All regressions include professor- and course-level fixed effects. Robust standard errors in parentheses. *** significant at 1%, ** at 5% and * at 10%.
Source: Authors' computation, using data from the University of Girona.

Table C.7: Results for off-campus students at the university's professional schools

	All students		Male students		Female students	
	Male lecturer (1)	Female lecturer (2)	Male lecturer (3)	Female lecturer (4)	Male lecturer (5)	Female lecturer (6)
Panel A: By treatment						
Treatment 1	0.012	0.000	-0.068	0.096	0.063	-0.061
	(0.04)	(0.05)	(0.06)	(0.08)	(0.06)	(0.06)
Treatment 2	0.040	0.050	-0.017	0.092	0.091	0.010
	(0.04)	(0.05)	(0.06)	(0.08)	(0.06)	(0.06)
Observations	3536	2993	1463	1041	2029	1900
R^2	0.270	0.314	0.304	0.377	0.307	0.374
Panel B: Intervention						
Treated	0.026	0.025	-0.042	0.094	0.077	-0.025
	(0.04)	(0.04)	(0.05)	(0.07)	(0.05)	(0.05)
Observations	3536	2993	1463	1041	2029	1900
R^2	0.270	0.314	0.304	0.377	0.306	0.373

Note: Each column in each panel is the result of a different regression, where the dependent variable is the teaching evaluation score given to lecturer j by student i relative to course s. All regressions include student characteristics (age, age-squared and final grade obtained) and professor- and course-level fixed effects. Columns (1) and (2) also control for students' gender. The number of observations in columns (3) and (5) may not add up to those of column (1) because the number of singleton observations in each regression may differ. The same is true in the case of columns (4) and (6), in comparison with column (2). Robust standard errors in parentheses. *** significant at 1%, ** at 5% and * at 10%.
Source: Authors' computation, using data from the University of Girona.

Table C.8: Results for the additional questions in the questionnaire

	All students		Male students		Female students	
	Male lecturer	Female lecturer	Male lecturer	Female lecturer	Male lecturer	Female lecturer
	(1)	(2)	(3)	(4)	(5)	(6)
Q5: The methods used by the professor help me to learn						
Treatment 1	-0.070***	-0.018	-0.048	-0.027	-0.083***	0.001
	(0.02)	(0.02)	(0.04)	(0.05)	(0.03)	(0.03)
Treatment 2	0.009	-0.014	0.021	-0.119**	0.006	0.034
	(0.02)	(0.02)	(0.04)	(0.05)	(0.03)	(0.03)
Observations	13281	12492	5641	3582	7324	8636
R2	0.368	0.363	0.407	0.392	0.409	0.398
Q6: I value positively the activities of support and office hours of the professor						
Treatment 1	-0.062***	-0.024	-0.035	-0.036	-0.076**	-0.013
	(0.02)	(0.02)	(0.04)	(0.05)	(0.03)	(0.03)
Treatment 2	-0.013	-0.020	-0.021	-0.127**	-0.015	0.020
	(0.02)	(0.02)	(0.04)	(0.05)	(0.03)	(0.03)
Observations	13281	12492	5641	3582	7324	8636
R2	0.350	0.341	0.379	0.376	0.398	0.375

Note: Each column in each panel is the result of a different regression, where the dependent variable is the teaching evaluation score given to lecturer j by student i relative to course s. All regressions include student characteristics (age, age-squared and final grade obtained) and professor- and course-level fixed effects. Columns (1) and (2) also control for students' gender. The number of observations in columns (3) and (5) may not add up to those of column (1) because the number of singleton observations in each regression may differ. The same is true in the case of columns (4) and (6), in comparison with column (2). Robust standard errors in parentheses. *** significant at 1%, ** at 5% and * at 10%.
Source: Authors' computation using data from the University of Girona.