Economics of Education Review 89 (2022) 102280

Contents lists available at ScienceDirect

Economics of Education Review

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

Online teaching and gender bias☆

Sara Ayllón *

Department of Economics, EQUALITAS & IZA University of Girona, Spain

ARTICLE INFO

JEL classification:
J16
J71
I23
J45

Keywords:
Gender bias
Online instruction
Teaching evaluations
Higher education
Spain

ABSTRACT

I study the impact of online instruction on teaching evaluations at a higher education institution in Spain. Using a difference-in-differences approach, I show that in the semester when teaching moved online, female lecturers were evaluated more poorly than in previous semesters. The performance of male lecturers was not impacted by the new teaching environment, according to student opinion. I rule out several mechanisms: for example, poorer adaptation to online teaching by female lecturers, less experience in taught courses or student sorting. Additional results indicate that among the female lecturers, those who were younger and who did not have a permanent contract were those impacted most negatively. The bias was driven by male students and by low achievers (who were going to fail the course), and was particularly pronounced in Social Sciences. If the online environment keeps gaining in importance in higher education, the gender gap in teaching evaluations that I document is likely to have important direct and indirect effects on the career progression of women.

1. Introduction

Online instruction in higher education has been gaining in importance in recent years. In the US, the percentage of students at Title IV institutions enrolled exclusively on distance-learning courses rose from 12.8% in 2013 to 17.3% in 2019, according to data from the National Center for Education Statistics. And among those students who were not exclusively engaged in distance learning, the percentage whose courses included some distance learning increased from 13.6% to 19.3% in the same period. Thus, even in 2019, close to four college students in 10 in the US received some instruction online. In Europe, the percentage of 16–24-year-olds who took a course online doubled in 2019 compared to 2010 (from 7% to 15%), according to data from Eurostat. And now, the coronavirus outbreak has greatly accelerated this process of change. In order to stop the spread of the coronavirus, universities all over the world moved teaching online. While many institutions are returning to face-to-face teaching, it is likely that online teaching will remain more common than was the case before the pandemic, and hybrid solutions are also likely to gain in relevance. Thus, it is important to consider the consequences of such an important change in the teaching environment at many different levels. While there has been a rapid expansion in the body of literature on the impact of online teaching in terms of student performance (Rodríguez-Planas, 2022a), mental well-being (Browning et al., 2021; Jaeger et al., 2021; Rodríguez-Planas, 2022b), career and earnings expectations (Aucejo et al., 2020; Jaeger et al., 2021) and evaluations of the college experience (Aucejo et al., 2021), much less attention has been paid to the impact of online instruction on teachers — and, more precisely, to the evaluation of online teaching and the extent to which it may have had a different impact on male and female instructors.[1]

The main objective of this study is to learn how online teaching may have affected the gender bias in teaching evaluation in higher education (Boring, 2017; Boring et al., 2016; Mengel et al., 2019). In the context of this paper, I refer to any gender differences in teaching evaluations that cannot be explained by lecturer performance, effectiveness or student sorting as gender bias. Do students evaluate their lecturers' performance more equally when teaching takes place virtually? Or, conversely, does online teaching contribute to a widening

☆ Support from REQ 2021 (*'Plan de Recuperación, Transformación y Resiliencia'*, Ministerio de Universidades and NextGenerationEU) and the projects PID2019-104619RB-C43 (funded by MCIN/AEI/10.13039/501100011033) and 2017-SGR-1571 is acknowledged. I want to thank the Department of Social Sciences at the University of Eastern Finland for its warm hospitality while writing this paper. Any errors or misinterpretations are my own.

* Correspondence to: C/Universtat de Girona 10. 17003, Girona, Spain.
E-mail address: sara.ayllon@udg.edu.

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

[2] According to MacNell et al. (2015), students expect male and female professors to behave differently: whereas men are supposed to be *effective* (professional, objective, authoritative and knowledgeable), women are expected to be *interpersonal* (warm, accessible, nurturing, supportive and personable). Students tend to be critical of female lecturers who do not behave as expected: women are sanctioned if they do not exhibit strong interpersonal traits, whereas men are not.

[3] Women have actually been found to experience more Zoom fatigue than men, because of non-verbal mechanisms ('mirror anxiety', feeling physically trapped, 'hyper gaze' from staring faces, etc.) (Fauville et al., 2021).

https://doi.org/10.1016/j.econedurev.2022.102280
Received 13 October 2021; Received in revised form 1 June 2022; Accepted 6 June 2022
Available online 27 June 2022

of the existing gender bias in teaching evaluation? Several mechanisms may be at play here. On the one hand, it could be that male and female instructors have adapted differently to online teaching. Women may have performed more poorly than their male counterparts because of the (already well-documented) grave difficulties they faced in reconciling work and childcare during the first waves of the pandemic (Adams-Prassl et al., 2020; Alon et al., 2020a, 2020b; Deryugina et al., 2021; Farré et al., 2022; Zamarro & Prados, 2021). In that case, the gender gap could widen as a result of a genuinely poorer performance by female instructors. On the other hand, online teaching may have narrowed the gender bias in teaching evaluation, if remote classes benefited from female teaching styles — which are thought to be more interpersonal (MacNell et al., 2015). Assuming that women are more likely to be supportive, accessible or personable than men, students may be more appreciative of the support received from their instructors in difficult times (such as during a pandemic) and that may be reflected in their evaluation of the teaching.[2] But again, if online instruction makes it more difficult for women to excel at this interpersonal teaching style — because of the difficulties in creating immediacy through verbal communication (given that non-verbal communication and body language are often eliminated) — they may be penalized.[3]

Furthermore, it could be that female instructors have less experience in the courses that – all of a sudden – need to be taught online; and again, that could be reflected in differences in the evaluations of the teaching of male and female lecturers. Also, one needs to consider the possibility that men and women teach subjects of a different nature, which could in turn have different degrees of adaptability to an online environment. Thus, it is important to discount the possibility that gender differences in teaching evaluations are not the result of self-selection (or sorting) by students into subjects. Nonetheless, if one can discount all the aforementioned mechanisms and still observe a gap in teaching evaluations to the detriment of women, it must be that online teaching contributes to the strengthening of gender bias (either because of prejudice or dislike, either conscious or not, either implicit or not) (Bertrand et al., 2005; Bohnet, 2016; Oreopoulos, 2011; Rooth, 2010).

A number of studies have previously investigated gender bias in teaching evaluation, with the results always pointing in the same direction: female instructors are discriminated against in evaluations of their teaching (Boring, 2017; Boring et al., 2016; Boring & Philippe, 2021; Mengel et al., 2019; Wagner et al., 2016).[4] These studies found that differences in the teaching scores by gender cannot be explained by differences in teacher effectiveness, performance or skills. However, none of the studies presents evidence of a sudden shock in the teaching environment such as that brought about by the pandemic. Furthermore, there is only one article (that I am aware of) that studies differences in teaching evaluations by gender in an online environment: MacNell et al. (2015) performed an analysis of gender bias in an online course using male and female avatars. The authors found that students rated perceived male instructors higher, regardless of the actual gender of the teacher.

This study has several strengths. First, it is not affected by reverse causality, as the teaching evaluation questionnaire was filled in by students before the final exams. Also, teachers only learn of their evaluations several weeks after term is over, and thus there is no possibility of retaliating against a poor evaluation. Secondly, the unique database that I work with contains important information regarding students' and lecturers' characteristics. All this information allows a more nuanced understanding of the mechanisms that drive the results. Thirdly, I have longitudinal data stretching back over several academic years, and so I can aim to cancel out potential underlying objective performance differences (such as teaching style or personality) via fixed effects by lecturer. Finally, my study includes the whole universe of teaching evaluations at the University of Girona (Spain), and so this is an institution-wide study, where all fields of learning are included and sample size is relatively large. Regarding the limitations of my analysis, it is important to take into account the fact that results are restricted to those students who voluntarily complete the teaching evaluation questionnaire. Yet, from a policy perspective, it is really only the opinions of those students that count when deciding to promote a lecturer or settling on his/her performance-related pay. These are the results that have real consequences — not those that could have been obtained had all students completed the questionnaire. Finally, beyond the final grade obtained, I have no information that could proxy effort by the students — for example, study hours.

My results are based on difference-in-differences estimates, which provide the within-teacher comparison of the evaluations before and after the switch to online teaching in different academic years. The main results indicate that women received poorer evaluations when instruction took place online than they did in previous semesters. The same does not hold true for male instructors: the new teaching environment had no impact on their teaching scores. Analysis of the potential mechanisms driving the results indicates that women did not perform any worse than their male counterparts, and nor were they less effective. I reach this conclusion by considering student opinion regarding the adaptation of the course to the online environment, other aspects of teacher performance (beyond the overall assessment), the students' final grades and the lecturers' experience in the subjects taught. Nor are the results driven by the self-selection of students onto courses. Subgroup analysis points strongly to gender bias: the results are particularly negative for females who are in a weaker position (younger and without a permanent contract), and they stem from male students and low achievers.

This paper contributes to several strands of literature. First, it provides additional evidence of the well-documented discrimination suffered by women in the labour market, tying in closely with studies on stereotyping and social constructs that contribute to such discrimination (Bagues & Esteve-Volart, 2010; Sarsons et al., 2021). Secondly, it adds to the growing body of literature on the short- and long-term effects of the coronavirus pandemic on females, and particularly on women in academia (Deryugina et al., 2021). It highlights the fact that both research and teaching need to be accounted for when considering the consequences of the COVID-19 outbreak on female academics. Thirdly, this paper swells the literature on the impact of new teaching environments on lecturers at the university level, given important concerns regarding increased levels of stress, burnout and teacher turnover since 2020 (Zamarro et al., 2022). And, finally, it makes a contribution to the ongoing discussion of the validity of teaching evaluations (Boring et al., 2016; Carrell & West, 2010; Hoffmann & Oreopoulos, 2009).

The paper is organized as follows. Section 2 provides the institutional background to my analysis regarding the outbreak of the coronavirus pandemic and the interventions taken by the University of Girona (which are similar to those taken by the great majority of universities worldwide). Section 3 presents the data and Section 4 the empirical strategy. Section 5 details the results, while Section 6 analyses the mechanisms driving my main findings. Section 7 provides a conclusion and discusses potential policy implications.

2. Institutional background

Spain was one of the epicentres of the coronavirus outbreak in Europe. It was hit early and hard. By 14 March 2020, the Spanish government recognized the need to declare a state of emergency. Following that, an immediate strict nationwide lockdown was mandated, which resulted in the closure of schools, universities and non-essential businesses. Everybody was ordered to remain at home, and only essential activities were allowed, such as buying food or medicines,

[4] Problems with teaching evaluations other than gender bias are detailed in Stark and Freishtat (2014). Other types of bias in teaching evaluation – for example, against non-native speakers or ethnic minorities – have been studied by Fan et al. (2019) and Wagner et al. (2016).

going to the doctor or caring for elderly persons. While going to work was permitted, working from home was encouraged. On 26 March and again on 9 April, the Spanish government extended the state of emergency. On 13 April, non-essential workers who could not work from home (e.g. in the construction sector or industrial production) were allowed to return to the workplace. On 22 April, the government – in the face of strong opposition – extended the state of emergency for another 15 days. Only on 27 April were children allowed to go outside (for the first time since mid-March). By the beginning of May, a de-escalation plan had started, under which restrictions were lifted region by region, in four phases, depending on the health conditions in the area. The state of emergency was finally lifted on 21 June, after more than three months of extraordinary restrictions due to the first wave of the coronavirus pandemic.

At the University of Girona, on 14 March, and immediately after the declaration of the state of emergency at the national level, the rector issued a resolution suspending all face-to-face educational activities — to include not only all teaching, but also seminars and workshops. Only certain activities deemed to be critical or essential were allowed — for example, laboratory maintenance or campus safety. According to the resolution, all teaching was to continue online, employing the digital tools available at the university. On 31 March, the rector issued a new resolution that simply extended the previous one. On 28 April, a further resolution was issued, by which research groups could apply for access to the university buildings, in order to do essential research tasks that could not be performed from home (e.g. in a laboratory). All teaching continued to be undertaken online. On 15 May, a new resolution was approved, extending the measures of the 31 March resolution. In May and June, final exams took place online and the academic year ended without seeing the return of students to classrooms. Altogether, in the second semester of the academic year 2019/20, students received face-to-face teaching for only about six weeks between the beginning of February and mid-March; the rest of the term was taught and evaluated completely online. Throughout the paper, I refer to this term as the 'online semester'.

3. Data

The data used for the empirical analysis is the whole universe of teaching evaluations by students at the University of Girona during the academic years 2018/19 and 2019/20.[5] Teaching is mainly organized in semesters, and thus the data covers four terms. Each semester, before the final exams, students are asked to fill in a questionnaire via the Moodle platform.[6] Students are not obliged to complete the teaching evaluation questionnaire, but they are encouraged to do so. They receive messages reminding them of the importance of filling in the questionnaire and are reassured that they cannot be identified or penalized in any way. Students can complete the questionnaire at any time of day during the three weeks that the questionnaire is active. All the answers are kept completely anonymous. Furthermore, the questionnaire is asked on all courses where a given lecturer teaches at least 1.5 European Credit Transfer and Accumulation System (ECTS) credits. Other aspects, such as the size of the class or the nature of the course (e.g. theoretical or practical), are not relevant. The same questionnaire is used across all faculties.

In total, I was provided with 76,346 complete observations; of these, 38,771 referred to academic year 2018/19 and 37,575 to 2019/20. In all, 1544 lecturers were evaluated in academic year 2018/19 and 1638 the following year. There were 1752 courses in 2018/19 and 1841 in 2019/20. A total of 5758 students filled in the questionnaire in academic year 2018/19 and 5426 in 2019/20. On average, each lecturer was evaluated by 25 students in academic year 2018/19 and by 23 in 2019/20.

Table 1
Teaching evaluation scores, summary statistics.
Source: Author's computation using data from teaching evaluations at the University of Girona, 2018/19, 2019/20.

Variable	Mean	Std. Dev.	N
Part B of the questionnaire:			
Teaching score (all semesters)	4.027	1.155	76 346
By semester and academic year:			
1st semester, 2018/19	4.011	1.136	23 148
2nd semester, 2018/19	4.013	1.153	15 623
1st semester, 2019/20	4.045	1.137	15 922
2nd semester, 2019/20 (online semester)	4.039	1.190	21 653
By lecturer's gender:			
Male	4.019	1.160	42 257
Female	4.036	1.150	34 089
By lecturer's age:			
44 or younger	4.144	1.103	26 971
45 or older	3.950	1.182	43 027
By lecturer's type of contract:			
Non-permanent	4.099	1.123	41 916
Permanent	3.938	1.188	34 430
Part A of the questionnaire:			
Statement #1	4.263	1.068	75 977
Statement #2	4.031	1.192	76 143
Statement #3	3.836	1.25	75 800
Statement #4	3.964	1.195	75 695
Statement #5	3.869	1.22	74 376
Statement #6	4.354	1.064	54 712

Note: All the possible values go from 1 ('strong disagreement') to 5 ('strong agreement'). The teaching score is calculated using the answers to the following statement: 'I evaluate this teacher's overall performance as positive'. As for the rest of statements: #1 'This teacher set out the course syllabus and the evaluation criteria clearly'; #2 'With this teacher, I learn'; #3 'This teacher motivates me to make an effort and to learn by myself'; #4 'The course material that the teacher provides me with helps'; #5 'The evaluation procedure allows me to demonstrate my knowledge'; and #6 'This teacher helped me overcome my doubts when I consulted him/her'. The unit of analysis is each student answer by academic year, semester, course and lecturer.

My main dependent variable contains the responses to the sole statement in Part B of the questionnaire: 'I evaluate this teacher's overall performance as positive'. This statement seeks to elicit an overall assessment by each student of the lecturer's work. Responses are on a Likert scale ranging from 1 to 5, where 1 indicates 'strong disagreement' and 5 'strong agreement'. The first panel of Table 1 shows that average evaluation in the period of analysis was 4.027, with a standard deviation of 1.155. The figure was very similar, regardless of the semester and academic year. The following rows in Table 1 also show that, on average, males and females were similarly evaluated, while younger lecturers (under 45 years of age) obtained higher scores than their older colleagues. In the same way, instructors on a temporary contract scored better than those on a permanent contract.[7]

In the analysis, I also use the responses to Part A of the questionnaire, which asks students their opinions on six different aspects of the lecturer's performance. The statements read as follows: (1) 'This teacher set out the course syllabus and the evaluation criteria clearly'; (2) 'With this teacher, I learn'; (3) 'This teacher motivates me to make an effort and to learn by myself'; (4) 'The course material that the teacher provides me with helps'; (5) 'The evaluation procedure allows me to demonstrate my knowledge'; and (6) 'This teacher helped me overcome my doubts when I consulted him/her'. The second panel of Table 1 details the mean and the standard deviation for each statement:

[5] The teaching evaluations from the academic year 2020/21 onwards cannot be used, as the questionnaire changed.

[6] When students complete the questionnaire, they may already know what their continuous assessment grade is (if any), but not the final grade that they will obtain on the course. Note that by 'final exams', I refer to those taken by students at the end of each term.

[7] Simple t-tests indicate that such differences are statistically significant.

Table 2
Sample characteristics.
Source: Author's computation using data from teaching evaluations at the University of Girona, 2018/19, 2019/20.

Variable	Mean	Std. Dev.	Min.	Max.
Students:				
Age	21.771	5.08	17.182	77.5
Female	0.582	0.493	0	1
Humanities	0.071	0.256	0	1
Sciences	0.115	0.319	0	1
Life Sciences	0.194	0.396	0	1
Social Sciences	0.466	0.499	0	1
Engineering	0.157	0.364	0	1
Final grade obtained	6.912	1.358	0.4	10
Course repeater	0.012	0.063	0	1
No. of observations — students:				8380
Lecturers:				
Age	45.880	10.095	23.333	73.368
Female	0.475	0.499	0	1
Permanent contract	0.281	0.449	0	1
No. of observations — lecturers:				1861

Note: The unit of analysis is each individual student (panel 1) and each individual lecturer (panel 2).

as can be seen, the mean for all statements hovers around the value of 4, with statements 1 (on the syllabus and evaluation criteria) and 6 (on teachers' approachability) gaining higher values, and statements 3 (on motivation) and 5 (on the evaluation procedure) receiving the lowest.[8] Over the course of the two academic years analysed and the four semesters, one can observe no great differences, with the values typically hovering around the period average (results not shown to save space, but available from the author on request). Finally, in the second semester of academic year 2019/20, three new statements were introduced to gather information on students' experiences during the online semester: (1) 'I am satisfied with the adaptation [to the online environment] of the course materials'; (2) 'The support activities and the tutorship of the lecturer during this period were satisfactory'; and, (3) 'The volume of work adapted [to the new online environment] has been coherent and proportionate to the number of course credits'. I call these the 'COVID statements' and I provide details on them in Section 6.1.1 below.

Table 2 shows sample characteristics. Regarding students (first panel), average age is 21.8 and 58% are females.[9] Nearly half of the sample was reading for a degree in Social Sciences (46.6%) and one in five was studying Life Sciences (19.4%). As for the remainder, the figures are Engineering — 15.7%; Sciences — 11.5%; and Humanities — 7.1%. Lastly, the final grade obtained can take any value from 0 to 10, to one decimal place, with a mean of 6.91 and a standard deviation of 1.358. About 1.2% of the students who filled in the questionnaire were retaking the course. Among lecturers (second panel), average age was nearly 46. About 47.5% of instructors were female and nearly three in 10 were on a permanent contract.

4. Empirical strategy

I base my results in difference-in-differences models, by which I compare evaluations for the same teacher in the second term of academic year 2019/20 (the online semester) to those in the first term of the same academic year, using the two terms of academic year 2018/19 as a control. Thus, my identification strategy relies on within-teacher comparison of the evaluations before and after the change to online teaching in different academic years.[10] My ultimate goal is to learn whether the online semester impacted male and female instructors differently, according to student opinion; and therefore, to learn whether online teaching served to widen the gender gap in teaching evaluations. Formally, Eq. (1) estimate is as follows:

$$E_{ilsa} = \alpha + \beta_1 \cdot S2 + \beta_2 \cdot a2019/20 + \beta_3 \cdot S2 \cdot a2019/20 + \gamma \cdot X_{isa} + L_l + \epsilon_{ilsa} \quad (1)$$

where i refers to a given student, l to a lecturer, s to a semester and a to an academic year. E_{ilsa} is the main outcome of interest and is the teaching evaluation provided by student i for lecturer l during semester s of academic year a. $S2$ is an indicator for the second term of the academic year (in both 2018/19 and 2019/20). $a2019/20$ is a dummy variable that takes the value 1 for the academic year 2019/20 and 0 otherwise. Such fixed effect by year controls for changes that may have occurred between academic years regarding, for example, university regulations. The coefficient of interest is β_3, which captures the average change in the outcome between the first and the second terms of 2019/20 above and beyond the existing difference between the first and the second terms of academic year 2018/19. X_{isa} are control variables that refer to the student (age, its square, gender, course repeater and field of study). L_l are fixed effects by lecturer to account for observed, observable or unobservable characteristics that are time invariant and may be relevant to a lecturer's performance (e.g. teaching style, personality, etc.). Thus, I identify from the within-lecturer variation. ϵ_{ilsa} is the usual error term. Standard errors are robust and clustered at the student level to account for the fact that the scores given by each student are not independent of one another.

Importantly, my identification strategy relies on the fact that student characteristics are similar, irrespective of whether students completed their teaching evaluations in the online semester or in face-to-face semesters. In Table A.1 in the Appendix, I test this hypothesis by running my main regression on students' observable characteristics, rather than the teaching evaluation scores. The results indicate that the online semester is not correlated with students' characteristics — with the important exception of grade and the probability that the student has been repeating a course. Average grades in the first and second semesters of 2018/19 and the first semester of 2019/20 are, respectively, 6.8, 7.0 and 6.8, while in the online semester it is 7.5. The upper left-hand graph in Fig. A.2 assesses the equality of the distribution function point by point for the two semesters of 2019/20.[11] The global test of equality of the two cumulative distribution functions (CDF) is rejected with a p-value of below 0.0001. Moreover, the CDF equality hypothesis is rejected at all points between 0.6 and 8.9, and at most points beyond the latter — see the thick horizontal line near the bottom of the graph, which shows the ranges where CDF equality is rejected. The grades are higher in the online semester than in the previous semester of the same academic year at nearly every point. The same is true if we compare the second semester of academic year 2018/19 and the online semester (upper right-hand graph). In the bottom left-hand and bottom right-hand graphs, one can check the striking similarity of grades in the first semester of both academic years analysed, and, to a lesser extent, also the two semesters within

[8] Fig. A.1 in the Appendix shows the frequency distribution of all scores to the teacher's overall performance evaluation and the rest of the statements in the questionnaire.

[9] Table 2 has been computed taking each individual student as the unit of analysis. The results are slightly different if I take each student–teacher–year observation as the unit of analysis. However, there is no difference worth mentioning. The same is true of the second panel, where I consider each individual lecturer as the unit of analysis.

[10] Ideally, I would be comparing within-teacher within-course between-term evaluations, but at the University of Girona the great majority of courses are taught in only one semester of the academic year.

[11] I use the Stata command *distcomp* for this purpose (Kaplan, 2019), which establishes whether and where two cumulative distribution functions are different, while accounting for skewness and scale differences.

academic year 2018/19. Thus, I discount the use of grade as a control in the main specifications.[12]

Furthermore, note also that column (3) in Table A.1 indicates that students who have been repeating the course are less likely to have completed the questionnaire in the online semester. Even when the coefficient is highly significant, it implies that the predicted probability of a student retaking a course in the online semester is 1.1%, compared to 1.5% in the remaining semesters. I consider the size of this effect to be economically insignificant and I include the variable as a control. Notably, all the main point estimates are virtually identical when the variable is excluded as a control. The same is true of the small changes in the percentage of students who completed the questionnaire, by field of study — see columns (5) to (9) in Table A.1.

5. Results

Table 3 presents the main results. The first column indicates that in the online semester, students evaluated their lecturers slightly worse than in the first semester of the same academic year and than in both semesters of 2018/19. The estimated coefficient is only statistically significant at 90However, when I use controls (student's age, its square, gender, whether the student is repeating the course, and field of study) and robust standard errors clustered at the student level (column 2), the coefficient is no longer statistically significant; this indicates that teaching evaluations during the online semester were, on average, no different from those in previous semesters. Interestingly, though, separate regressions by gender of the lecturer indicate a different story. Columns (3) and (4) show that the online semester had, on average, no impact on the evaluation of male lecturers; but for female lecturers, the average evaluation score decreased by 0.063 points in the online semester compared to previous semesters (about 5.4% of a standard deviation). Thus, while the new teaching environment had, on average, no effect on men's scores, it did negatively impact the scores received by women.[13]

Previous results can be qualified with a number of exercises. First, I ensured that the results do not depend on a comparison between academic years 2018/19 and 2019/20 exclusively, by including data from academic year 2015/16 onwards in the sample. When I did that, the main findings remained the same, although both the negative coefficient for the online semester among women and its statistical significance became smaller.[14] Secondly, I confirmed that academic year 2018/19 was a good 'control' year, by running all the specifications for a sample that included data from only academic years 2017/18 and 2018/19, and by simulating a scenario whereby the online semester occurred in the second semester of 2018/19. The results of this placebo exercise indicated a non-negative coefficient for the teaching evaluations of female lecturers in the second semester of 2018/19, compared to the first semester of the same academic year and both semesters of academic year 2017/18. Finally, I also added fixed effects by degree (and dropped the control for field of study) to account for the possibility that different degree programmes can present greater or lesser difficulty in adapting to the new online environment. The results remained the same when I did that.[15] All these results are detailed in Table A.2 in Appendix.

In the next section, we learn about the mechanisms underlying the different impact of online teaching on the scores received by male and female instructors.

Table 3
Difference-in-differences results for teaching evaluations at the University of Girona, all lecturers and by lecturer's gender.
Source: Author's computation using the whole universe of teaching evaluations at the University of Girona, 2018/19 and 2019/20.

	All lecturers		By gender	
			Male	Female
	(1)	(2)	(3)	(4)
Online semester	−0.0294*	−0.0301	−0.0065	−0.0626**
	(0.0163)	(0.0210)	(0.0257)	(0.0293)
Academic year 2019/20	0.0339***	0.0336**	0.0184	0.0547***
	(0.0113)	(0.0142)	(0.0169)	(0.0203)
2nd semester	0.0033	−0.0141	−0.0497**	0.0326
	(0.0131)	(0.0163)	(0.0203)	(0.0227)
Prof. FE	Yes	Yes	Yes	Yes
Controls	No	Yes	Yes	Yes
Robust std. errors	No	Yes	Yes	Yes
Clustered std. errors	No	Yes	Yes	Yes
Observations	75 822	75 822	41 986	33 836

Note: Controls include student age, its square, student gender, course repeater and field of study. Significance level: ***p < 0.01, **p < 0.05, *p < 0.1.

6. Mechanisms

Are the gendered differences in teaching evaluations during the online semester the result of poorer performance by female instructors or student sorting, or, alternatively, is it gender bias? In order to answer this question, first I assess how male and female instructors actually performed during the online semester, using different proxies: (1) I analyse whether there are gender differences in terms of the adaptation of lecturers to the new teaching environment; (2) I look at whether the final grade – which can be thought of as an objective measure of the instructor's effectiveness – can explain the gendered differences in teaching evaluations; (3) I investigate whether differences in the experience of male and female instructors could help understand the gender gap in scores; and (4) I evaluate other aspects of the teacher's performance, as seen by the students (for example, support materials and the evaluation procedure). Secondly, I consider the possibility that the gendered differences in teaching evaluations may be a result of sorting by students, who may self-select onto certain courses (disproportionately taught by a given gender). And, finally, I break down the results by subgroup – considering lecturers' characteristics, students' characteristics and field of study – to gain a more nuanced understanding of who is driving the results.

6.1. Did female instructors perform more poorly than male instructors during the online semester?

6.1.1. Teaching in a new environment

The gendered results found in the previous section could be a consequence of female lecturers having greater difficulty in adapting

[12] Note that I also opt not to include information on grades from previous academic years because of sample loss, and also because I only know the final grade obtained in courses where the student filled in the questionnaire, and so it is not necessarily a good proxy for achievement or effort.

[13] While summary statistics in Table 1 indicate that males and females were, on average, similarly evaluated during the period of analysis, an OLS regression of the teaching evaluation scores against a dummy that indicates whether the lecturer is male or female and also controls for a number of important characteristics, points to a gender gap of −0.053 (statistically significant at 99%) for the three semesters prior to the pandemic. In the specification, I control for lecturer age and its square, lecturer type of contract, spring semester, mandatory course, field of study, academic year, student age and its square, student gender, final grade and course repeater; standard errors are clustered at the student level. That is, the gender gap in teaching evaluations at the University of Girona existed before the online semester, though it was possibly smaller than in other contexts — 4.7% of a standard deviation compared to, for example, 20.7% in Mengel et al. (2019), with all the caveats that such comparison entails.

[14] Note that for my main results I chose to work with the sample for academic year 2018/19 because it was the closest (and therefore most comparable) to the sample for 2019/20.

[15] Note that this is not my preferred specification, since certain degrees or special programmes have very few observations.

Table 4
Mechanism checks — Did female lecturers perform more poorly during the online semester?
Source: Author's computation using data from teaching evaluations at the University of Girona, 2nd semester, 2019/20.

	COVID statement #1 (1)	COVID statement #2 (2)	COVID statement #3 (3)	Final grade (4)	Final grade (shared subjects) (5)	Final grade (6)	Final grade (shared subjects) (7)	Previously taught course (8)
Female lecturer	−0.0243	−0.0137	−0.0171	0.1219***	0.0937***			0.0130**
	(0.0225)	(0.0221)	(0.0213)	(0.0219)	(0.0242)			(0.0065)
Ref. Male lecturer, male student								
Male lecturer, female student						0.1987***	0.2662***	
						(0.0581)	(0.0759)	
Female lecturer, male student						0.0475	0.0601	
						(0.0444)	(0.0460)	
Female lecturer, female student						0.3565***	0.3763***	
						(0.0565)	(0.0735)	
Observations	16 938	16 675	16 959	19 915	10 685	19 915	10 685	19 984

Note: The dependent variable of each regression is detailed in the column header. Regressions in columns (1) to (7) include lecturer's age (and its square), whether the lecturer holds a permanent contract or not, student gender, student age (and its square), final grade, course repeater and field of study. Standard errors are clustered at the student level. Shared courses in columns (5) and (7) refer to courses taught by both males and females. Column (8) only includes controls at the lecturer level and field of study. Significance level: ***p < 0.01, **p < 0.05, *p < 0.1.

their teaching to online classes.[16] If that were the case, it could explain why students judge the performance of their female lecturers to have been worse during the online semester. To investigate this possibility, I take advantage of three additional statements (the so-called 'COVID statements') that were asked of students in the evaluation questionnaire during the online semester. Again, on a Likert scale from 1 to 5, students expressed the degree to which they disagreed or agreed with the following statements:

1. 'I am satisfied with the adaptation [to the online environment] of the course materials'.
2. 'The support activities and the tutorship of the lecturer during this period were satisfactory'.
3. 'The volume of work adapted [to the new online environment] has been coherent and proportionate to the number of course credits'.

Fig. 1 shows that students were equally satisfied with the adaptation of the course materials to online teaching by male and female lecturers. They also believed they had received a similar level of support from their male and their female lecturers. And a similar number of them considered that the amount of work assigned during the semester was consistent and proportionate, regardless of whether their lecturer was a woman or a man. The first three columns of Table 4 also show the results of ordinary least squares (OLS) regressions, where I regress the gender of a lecturer against student responses to the COVID statements. Controls include the instructor's age (and its square), whether or not the lecturer is on a permanent contract, student gender, student age (and its square), final grade, whether the student has been repeating the course, and field of study.[17] Standard errors are clustered at the student level. In all regressions, the coefficient for female lecturer is not statistically significant. Thus, male and female lecturers did not do a markedly different job of adapting to virtual teaching, according to student opinion, and so one can disregard the possibility that the results are explained by a belief among students that their female lecturers were genuinely worse when they had to teach in a new online environment.

[16] There could be various reasons, but possibly the most important is related to the difficulties faced by women in reconciling work and childcare since the coronavirus outbreak (Adams-Prassl et al., 2020; Alon et al., 2020a, 2020b; Farré et al., 2022; Zamarro & Prados, 2021). See also Deryugina et al. (2021) for evidence on female academics.

[17] The results in this section and the next one only apply to the online semester, and so the use of the information on the final grade is not subject to the comparison problems commented on in Section 4.

6.1.2. Final grade

Assuming that a student's final grade is an objective measure of an instructor's performance, we would expect the effect of female teachers on grades to be negative and statistically significant if women were genuinely worse than men as lecturers during the online semester.[18] Column (4) in Table 4 indicates that this is not the case. A simple regression of a lecturer's gender against students' final grades in the online semester indicates that, if anything, the students of female instructors obtained higher grades. Naturally, it could be hypothesized that women perhaps tried to compensate for their poorer performance by giving students better grades. Yet I find that this is not the case. In an additional specification, where I consider only the sample of observations belonging to courses taught by both men and women (column 5), with typically the same final exam, I continue to find that female instructors were more effective.

Below, we discover that the bias against women is mostly driven by male students, who give worse scores to female lecturers. Accordingly, I have also checked here the possibility that the results could be driven by female lecturers giving male students worse grades in the online semester than they received from male lecturers. I run a regression with a series of dummies that combine student and lecturer gender against students' final grades. I find that male students obtained lower grades than did female students, but the final grade awarded was no different depending on whether the lecturer was a man or a woman (column 6). The same is true in the sample of courses taught by both men and women (column 7).[19] Thus, the reason why students gave lower scores to female instructors in the teaching evaluations was not related to inferior learning outcomes. Differences in teaching skills by gender did not drive gender differences in the evaluation of instructors.

6.1.3. Teaching experience

There is a possibility that the results could have been driven by the fact that, as they sometimes teach different courses in different semesters, lecturers could have performed worse if, for example, they were teaching a particular course for the first time during the online semester. Lack of experience in the course would add to their lack of experience of teaching it in an online environment. If this was disproportionately true of female lecturers, the gendered results presented in Section 5 could be explained by lack of experience, rather than

[18] Recall that students do not know their final grades when they fill in the teaching evaluation questionnaire.

[19] The results also indicate that female lecturers gave female students better grades in the online semester than did male lecturers; however, as we discover below, that did not have any effect on female teaching evaluations.

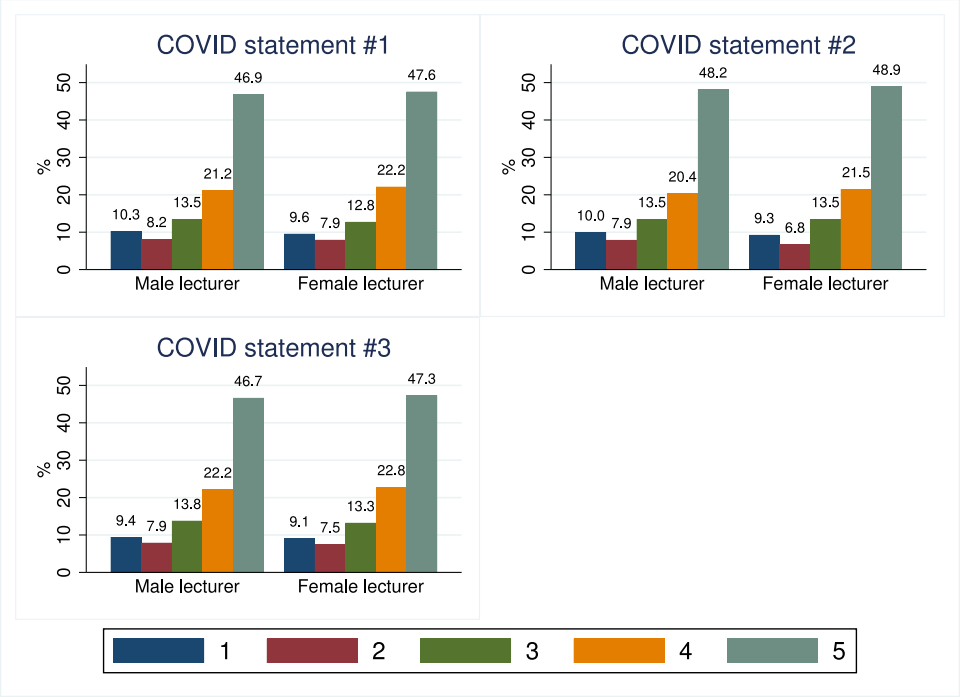

Fig. 1. Percentage of responses to the COVID statements by lecturer's gender, 2nd semester, 2019/20 (online semester). Note: In the graphs, 1 indicates 'strong disagreement' while 5 indicates 'strong agreement'. COVID statement #1 reads 'I am satisfied with the adaptation [to the online environment] of the course materials'. COVID statement #2 reads 'The support activities and the tutorship of the lecturer during this period were satisfactory'. And COVID statement #3 reads 'The volume of work adapted [to the new online environment] has been coherent and proportionate to the number of course credits'.
Source: Author's computation using data from teaching evaluations at the University of Girona, 2nd semester, 2019/20.

anything else. But the descriptive statistics indicate that this was not the case — see Fig. 2. Among male lecturers, about 60.5% of the courses during the online semester had already been taught in the previous academic year by the same person. In the case of females, the percentage was 58.2%. A simple regression for the likelihood of the same course being taught as in the previous academic year against a lecturer's gender (together with controls at the instructor level) indicates that during the online semester women were more likely to have taught a course that they had taught the previous academic year (see column (8) in Table 4). Thus, the results cannot be explained by female lecturers disproportionately teaching courses for which they lacked experience in the online semester.

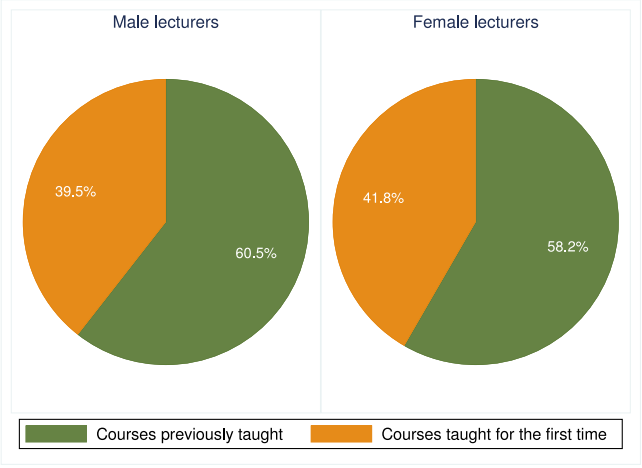

Fig. 2. Percentage of courses taught for the first time and previously taught courses by lecturer's gender, 2nd semester of 2019/20 (online semester).
Source: Author's computation using data from teaching evaluations at the University of Girona, 2018/19, 2019/20.

6.1.4. *Other aspects of lecturer performance*

Finally, I consider the remaining questions in the teaching evaluation questionnaire (Part A) that gather students' opinions on six different aspects of a lecturer's performance: (1) presentation of the course syllabus and the evaluation criteria, (2) the extent to which students feel they are learning with their instructor, (3) whether they are motivated to make an effort and learn by themselves, (4) the quality of the support materials, (5) the evaluation procedure, and (6) whether they obtained help from the instructor if they sought it. Interestingly, separate regressions by gender (not shown) for each of these outcomes yielded no significant results, indicating that the online semester did not have any impact on the scores received by male and female instructors when other aspects of the lecturer's performance, beyond the overall assessment, were evaluated. Thus, the gendered difference in the teaching evaluation result of the online environment does not appear to be driven by (potentially more objective) aspects of the teacher's performance. The bias creeps in when students evaluate overall performance.[20]

Taken together, the results suggest that the poorer evaluation of female lecturers in the online semester stems neither from objective

[20] In a departure from Mengel et al. (2019), it would seem that students do not base their responses about the lecturer's overall assessment on their previous responses to specific aspects of the lecturer's performance.

Table 5
Mechanism checks — Is student sorting into courses driving the results?
Source: Author's computation using the whole universe of teaching evaluations at the University of Girona, 2018/19 and 2019/20.

	Mandatory courses		Courses taught by males and females	
	Male	Female	Male	Female
	(1)	(2)	(3)	(4)
Online semester	0.0065	−0.0871***	0.0739*	−0.1325***
	(0.0271)	(0.0314)	(0.0408)	(0.0418)
Academic year 2019/20	0.0060	0.0796***	−0.0499*	0.0503*
	(0.0177)	(0.0215)	(0.0288)	(0.0298)
2nd semester	−0.0662***	0.0208	−0.1506***	0.1094***
	(0.0218)	(0.0254)	(0.0346)	(0.0365)
Observations	37 752	29 669	18 409	16 555

Note: Controls include student age, its square, student gender, course repeater and field of study. Fixed effects by instructor and academic year are also included. Standard errors are robust and clustered at the student level. Significance level: ***p < 0.01, **p < 0.05, *p < 0.1.

differences in their teaching skills (measured by grades or previous experience in the courses taught), nor from a subjective opinion of the lecturer's effectiveness provided by the students themselves either relative to the actual online performance or with regard to other aspects of teaching. The results indicate bias against female lecturers that is not related to teaching quality.

6.2. What about student sorting?

The gendered results presented above are not driven by poorer performance among female lecturers in the online semester, but they could be driven by the self-selection of students onto courses (based on the instructor's gender).[21] To discount this possibility, I run the main specification for the sample of courses that are mandatory — in this way, I avoid any potential bias due to sorting. It is worth noting that in the first two years of a degree at the University of Girona, students register for a fixed combination of mandatory courses that simply defines on what day of the week a given course is delivered.[22] In their third and fourth years, students design their own timetables, combining mandatory and elective courses. Yet, given the smaller number of students at this level of specialization, mandatory courses typically have only one group, and thus students have no choice. Columns (1) and (2) of Table 5 show the main results for the sample of mandatory courses. Female lecturers received a poorer evaluation in the online semester (compared to previous semesters) of about 0.087 points, which is the equivalent of 7.6% of a standard deviation. The evaluation of male instructors remained unaffected during the online semester. Importantly, the results highlight the fact that my main findings are not driven by sorting.

Next, I take into account the possibility that students assess female instructors more poorly because women are more likely than men to, for example, teach courses (whether mandatory or not) that are more difficult or less easily adapted to online instruction. If that were the case, the teaching evaluations could reflect the degree of difficulty or the nature of a given course, rather than anything else. To ensure that my main findings are not driven by this possibility, I run the main specification for the sample of courses that are taught by both male and female instructors. The results in columns (3) and (4) of Table 5 indicate that, in this sample, the gender gap facing women is even greater (0.13 points), and that indeed students discriminate in favour of men. Thus, it would seem that when students could compare men and women within a particular course, they penalized the women even more during the online semester.[23]

Finally, another source of concern is that the main findings could be driven by certain student characteristics potentially correlated with instructor gender. A simple regression (not shown) for the online semester of lecturer gender against all the student observed characteristics, together with fixed effects by course – to account for the different nature of the courses taught – indicates that instructor gender is not correlated with student characteristics when teaching takes place online.

The results of the three exercises in this section all point in the same direction. The poorer evaluations of female lecturers received during the online semester (compared to previous semesters) are not the result of sorting by students.

6.3. Who is driving the results?

Given that the gendered results presented above cannot be rationalized by poorer performance on the part of female lecturers or sorting by students, it is important to learn which lecturers, students or fields of study are driving the results. In what follows, I run Eq. (1) by subgroup to obtain a more nuanced understanding of the mechanisms behind the main findings.

6.3.1. Lecturers

In order to investigate the profile of lecturers who were particularly affected by poorer evaluations in the online semester, I run separate regressions that not only consider the gender of teachers, but also their age and the type of contract that they hold. The first two columns of Panel A in Table 6 show the results for male and female lecturers aged under 45, while columns (3) and (4) detail the results for those aged 45 or older. Interestingly, the results for women indicate that only those under 45 performed more poorly in the online semester, according to student opinion. The associated coefficient is −0.10 (statistically significant at 95%). This is equivalent to 8.9% of a standard deviation worse evaluation for females in the online semester than in previous terms. On the other hand, students believed that male teachers under 45 performed better in the online semester than in previous semesters. Their evaluation was about 0.09 points above the average for the other semesters. As a result, in the online semester the differences between male and female lecturers in the same age group and at the same level of seniority are very large. The results are also negative in the case of older instructors, but the coefficient is not statistically significant in the case of women, while it is in the case of men. Given that the coefficients for younger and older male instructors are similar in magnitude, but of different sign, they cancel each other out; that is why, on average, we observe no change in the online semester for the male sample.

[21] Note that if self-selection is at play, it cannot be attributed to the pandemic, since students chose all their courses at the beginning of the academic year, and the online semester started months after the students had chosen their courses.

[22] For example, students belonging to Group A are taught Microeconomics on Mondays, while students belonging to Group B are taught the same course on Tuesdays. All groups pursuing the same degree are taught the same subjects.

[23] These results are in line with those of Wagner et al. (2016), with data from a Dutch university.

Table 6
Difference-in-differences results for teaching evaluations by lecturer's gender while considering lecturer's age, lecturer's type of contract, student gender and student final grades.
Source: Author's computation using the whole universe of teaching evaluations at the University of Girona, 2018/19 and 2019/20.

Panel A:	By lecturer's age			
	Younger than 45		45 or more	
	Male (1)	Female (2)	Male (3)	Female (4)
Online semester	0.0986**	−0.1033**	−0.0760**	−0.0580
	(0.0409)	(0.0444)	(0.0321)	(0.0377)
Academic year 2019/20	−0.0183	0.0796**	0.0442**	0.0313
	(0.0261)	(0.0309)	(0.0213)	(0.0264)
2nd semester	−0.1773***	0.0554	−0.0036	0.0607**
	(0.0339)	(0.0349)	(0.0253)	(0.0299)
Observations	13 428	13 364	25 406	17 288
Panel B:	By lecturer's type of contract			
	Non-permanent		Permanent	
	Male	Female	Male	Female
Online semester	0.0516	−0.0997***	−0.0646*	−0.0165
	(0.0329)	(0.0371)	(0.0373)	(0.0422)
Academic year 2019/20	−0.0048	0.0769***	0.0405*	0.0318
	(0.0219)	(0.0280)	(0.0238)	(0.0273)
2nd semester	−0.0847***	0.0306	−0.0180	0.0359
	(0.0272)	(0.0296)	(0.0287)	(0.0323)
Observations	22 332	19 288	19 654	14 548
Panel C:	By student gender			
	Male students		Female students	
	Male	Female	Male	Female
Online semester	−0.0502	−0.1518***	0.0262	−0.0097
	(0.0396)	(0.0534)	(0.0345)	(0.0357)
Academic year 2019/20	0.0330	0.0461	0.0048	0.0515**
	(0.0258)	(0.0333)	(0.0225)	(0.0258)
2nd semester	−0.0501*	0.0512	−0.0577**	0.0102
	(0.0300)	(0.0423)	(0.0283)	(0.0273)
Observations	19 175	11 303	22 712	22 400
Panel D:	By student final grades			
	Low achievers (<5)		High achievers (≥9)	
	Male	Female	Male	Female
Online semester	−0.0902	−0.3453**	0.1540**	−0.0618
	(0.1293)	(0.1537)	(0.0695)	(0.0712)
Academic year 2019/20	−0.0385	−0.0057	−0.0518	0.0550
	(0.0675)	(0.0901)	(0.0508)	(0.0551)
2nd semester	−0.1194	0.0417	−0.0533	0.0948
	(0.0966)	(0.1107)	(0.0560)	(0.0588)
Observations	2714	1775	5258	4631

Note. Controls include student age, its square, student gender, course repeater and field of study. All regressions include lecturer and year fixed effects. Standard errors are clustered at the student level. Significance level: ***p < 0.01, **p < 0.05, *p < 0.1.

Panel B of Table 6 confirms the previous findings when considering separate regressions by lecturer gender and type of contract (permanent or not) that each instructor holds. The results show that female lecturers without a permanent contract (and typically younger) were perceived by students to have performed more poorly.[24] The online semester left the evaluation of female lecturers on a permanent contract unaffected. In the case of males, again, I obtain a negative coefficient in the case of lecturers with a permanent contract (and typically older) — though statistical significance is below 95%.

In short, the student bias is stronger against junior women without a permanent contract. It would seem that seniority prevented more established female lecturers from being further penalized by students in the teaching evaluations during the online semester.

6.3.2. Students

In what follows, I assess the extent to which the results may differ by student gender, as previously documented in the literature (Boring, 2017; Boring et al., 2016; Boring & Philippe, 2021; Fan et al., 2019; Mengel et al., 2019). Panel C in Table 6 presents the results for separate regressions that considered both the gender of the lecturer and that of the student. Interestingly, male students were the only ones to consider that female lecturers had done their job more poorly during the online semester than in previous semesters. The coefficient is one of the largest found (−0.15 points) and it is statistically significant at 99%. Male students did not rate the performance of their male teachers as worse (or indeed as better). Importantly, female students believed that the online semester had no impact on the teaching performance of either their male or their female lecturers. While online teaching did not lead to any positive bias among female students, it did exacerbate discrimination against female lecturers by male students.[25]

Another piece of information that may help us gain a more exact profile of who the students were who drove the main findings is the final grade obtained by the student on each course. Importantly, grades were not used as a control variable in the regressions, because their distribution in the online semester was very different from previous semesters (as explained in Section 4). For the same reason, these results need to be treated with caution. I ran separate regressions for 'low achievers' (i.e. students with grades below 5, and who therefore failed the course in a given semester of a given academic year) and for 'high achievers', whose grades were 9 or above (10 is the highest possible mark at the University of Girona) and therefore obtained 'Excellent' or 'With honours'. I found that the gender bias against women was, to a large extent, explained by students who were set to fail the course: they assessed the ability of their female teachers more poorly during the online semester, but were neutral about the ability of their male lecturers, even if they were set to fail that course as well.[26] High achievers were more generous in their assessments of their male instructors, but not their female lecturers. Thus, the poorer evaluation of female lecturers by low achievers is not compensated for by a better evaluation from high achievers; meanwhile, high achievers do discriminate in favour of their male lecturers. All in all, women lose out.

6.3.3. Field of study

Next, I consider whether gender bias stems from particularly gender-imbalanced fields of study. The results in Table 7 indicate that of the five main fields of study, bias against women in the online semester was most likely to have occurred in Social Sciences. Note the negative coefficient in column (8) (statistically significant at 95%). To a lesser degree, women lecturers were also regarded as having performed more poorly in Sciences, though that coefficient is statistically significant only at 90%. Students of Humanities, Life Sciences and Engineering did not perceive the performance of their female lecturers to have been worse (or indeed better) during the online semester. The same was true of male teachers in Sciences, Life Sciences and Social Sciences. Note, however, the negative coefficients among male teachers in Humanities and Engineering. Unfortunately, there is no further information available that would allow me to gain a deeper understanding of the potential mechanisms behind these results. Interestingly enough, in

[24] Female lecturers without a permanent contract are, on average, 41 years of age, while male instructors on the same type of contract are, on average, 43. Male and female lecturers on a permanent contract are, on average 54 and 52, respectively.

[25] See Boring and Philippe (2021) for another analysis that confirms lack of positive bias among girls.

[26] Boring (2017) also discusses the fact that students apply double standards: they tend to be harsher toward female lecturers (but not toward male lecturers) if they receive bad grades. See also Sinclair and Kunda (2000).

Table 7
Difference-in-differences results for teaching evaluations at the University of Girona by lecturer's gender and field of study.
Source: Author's computation using the whole universe of teaching evaluations at the University of Girona, 2018/19 and 2019/20.

	Humanities		Sciences		Life Sciences		Social Sciences		Engineering	
	Male (1)	Female (2)	Male (3)	Female (4)	Male (5)	Female (6)	Male (7)	Female (8)	Male (9)	Female (10)
Online semester	−0.2666***	0.1290	0.0626	−0.1225*	0.0422	0.0381	0.0371	−0.0872**	−0.1168**	−0.0982
	(0.1027)	(0.1393)	(0.0720)	(0.0721)	(0.0594)	(0.0625)	(0.0413)	(0.0433)	(0.0536)	(0.0849)
Academic year 2019/20	0.1894**	−0.1828	−0.0148	0.0085	−0.0889**	0.1320***	−0.0165	0.0547*	0.1176***	0.0790
	(0.0755)	(0.1179)	(0.0453)	(0.0480)	(0.0350)	(0.0437)	(0.0262)	(0.0306)	(0.0381)	(0.0522)
2nd semester	0.0285	0.0303	−0.0764	−0.0124	0.0135	−0.0902*	−0.0599*	0.0637*	−0.0435	0.1357**
	(0.0804)	(0.1020)	(0.0586)	(0.0512)	(0.0511)	(0.0546)	(0.0337)	(0.0348)	(0.0405)	(0.0627)
Observations	2531	1638	5795	6157	6789	7779	16 821	14 858	10 041	3393

Note: Controls include student age, its square, student gender, course repeater and field of study. All regressions include lecturer and year fixed effects. Standard errors are clustered at the student level. Significance level: ***p < 0.01, **p < 0.05, *p < 0.1.

Humanities only 39% of all the assessments related to female lecturers, and in Engineering — just 20%; in the remaining three fields of study, the figure was close to 50% or above. Importantly, the results indicate that my findings are not driven by minority faculty teaching in fields of study where females make up a smaller percentage of the faculty.[27]

7. Concluding remarks

This paper presents evidence that online teaching in higher education – which has witnessed a massive surge since the coronavirus outbreak – has had a negative impact on the evaluation of the work of female lecturers. Results from difference-in-differences models that compare the teaching scores obtained in the online semester (the second semester of academic year 2019/20) against those of the first semester of the same academic year and both semesters of the previous academic year (2018/19) indicate that, on average, women received 0.063 fewer points (about 5.4% of a standard deviation) than in previous evaluations. By contrast, according to student opinion, the teaching performance of male lecturers remained, on average, unaffected in the new teaching environment.

Analysis of potential mechanisms indicates that these results are not a consequence of poorer performance by female lecturers during the online semester. Students themselves confirm this in additional statements added to the evaluation questionnaire during the online semester that refer to the adaptation of support materials to the new teaching environment, the support and tutoring activities of the teacher and the amount of work the students were expected to fulfil during the online semester. According to the subjective opinion of students, female instructors did not do a markedly worse job than their male colleagues during the online semester. Analysis of an objective measure of lecturer effectiveness – the final grade obtained by a student – also indicates that it is not differences in teaching skills that drive gender differences in evaluations. Nor was lack of experience in the subjects taught during the online semester the source of bias in the teaching evaluations. Interestingly, additional analysis of other aspects of a lecturer's performance – the course syllabus, the evaluation criteria, the quality of the support materials and the evaluation procedure – indicates that bias only creeps in when students evaluate the overall performance of a teacher, not specific dimensions.

Additional checks also indicate that the main findings were not driven by the sorting of students onto courses. The gendered results were confirmed for the sample of mandatory courses onto which students could not self-select. The same was true for the sample of courses that were taught by both male and female lecturers, thus discounting the possibility that the gender bias found was a consequence of different degrees of difficulty or adaptability to virtual teaching of the courses taught by male and female instructors. Nor did I find student characteristics that were strongly correlated with the gender of a lecturer in the online semester.

Given that neither lecturer performance and effectiveness nor student sorting can explain the main results, I attribute my gendered findings to bias. Subgroup analysis confirms this possibility, as the results are particularly negative for young female instructors without a permanent contract, and are strongly driven by male students and low achievers who – even before they know their final grade – retaliate against female instructors, but not against male teachers. The findings are most apparent in Social Sciences. Online teaching did not lead to any positive bias on the part of female students towards female instructors. Yet a considerable degree of discrimination in favour of male instructors is found among high-achieving students.

The ultimate reason why students awarded lower scores to women in the online semester than in previous semesters is difficult to discern. It could be that online teaching does not allow teachers to fulfil students' gendered expectations. Students have higher interpersonal expectations of their female instructors, and so if online teaching prevents female lecturers from being as supportive and personable as in face-to-face teaching, that may translate into a greater burden for female instructors. Students may unconsciously be displaying such bias (Bohnet, 2016).

The consequences of the results presented in this paper are multiple and make themselves felt at many different levels. Teaching evaluations are still used in hiring, firing and promotion decisions at many universities around the world. They are taken into account in granting tenure and in performance-related pay; in teaching awards; and even in future course selection. Teaching evaluations can have an impact on academics' mental health and well-being (Fan et al., 2019; Henning et al., 2018). Women may lose confidence in their teaching ability after receiving a poor evaluation. As a result, they may invest more time in preparing their courses than in their research, which could in turn have consequences for the progression of their careers. If the results presented in this paper regarding the impact of online teaching on female instructors' evaluations are validated in other contexts, the use of teaching assessments in charting women's careers may become more open to question, given that online teaching is gaining in importance in higher education.

The results presented in this paper have some important policy implications. This study calls for a reassessment of the current system for evaluating the quality of an instructor in higher education via

[27] Mengel et al. (2019) also find that gender bias is independent of whether the majority of instructors on a course are male or female. In their words: '[...] the bias we identify is a bias against female instructors per se rather than a bias against minority faculty in gender-imbalanced areas' (Mengel et al., 2019: 27).

teaching scores, as it systematically disadvantages women in academia. Evaluations should not serve as a means of benefiting or further enhancing the position of those overrepresented in the upper echelons of university leadership (McElroy, 2016). If – regardless of the results presented here – teaching evaluations are seen as a useful tool, then there needs to be progress to eradicate (implicit or explicit) bias against female instructors. Boring and Philippe (2021) provide an excellent example of a successful intervention, whereby emails were sent out to students to make them aware of the existing bias.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

A.1. Figures

See Figs. A.1 and A.2.

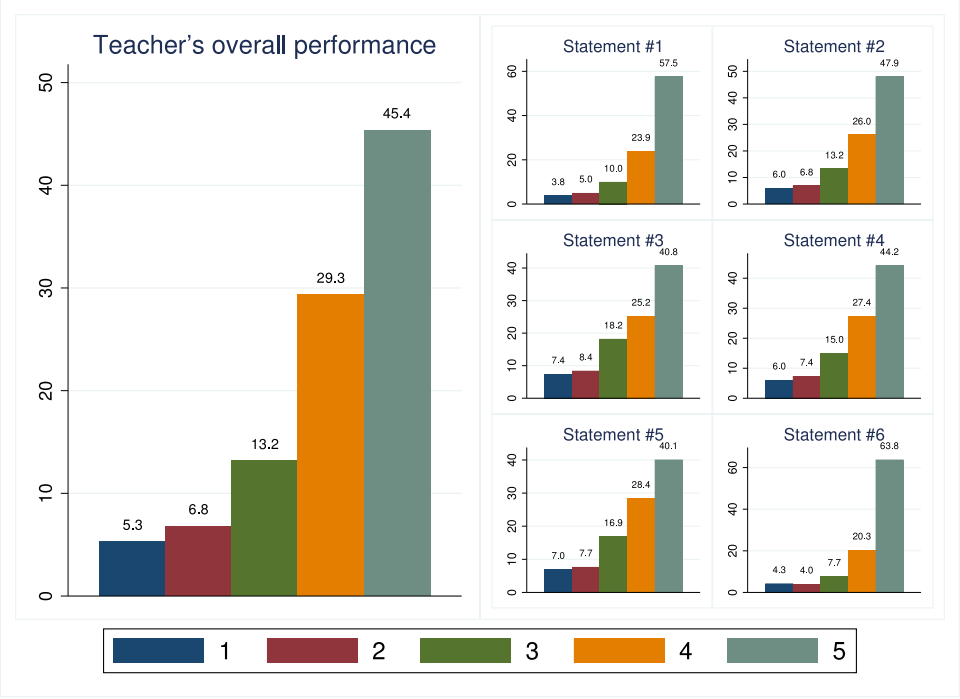

Fig. A.1. Frequency distribution of scores for the teacher's overall performance evaluation (left) and the rest of the statements in the teaching questionnaire (right), 2018/19 and 2019/20. Note: In the graphs, 1 indicates 'strong disagreement', while 5 indicates 'strong agreement'. The teaching score statement reads as 'I evaluate this teacher's overall performance as positive'. Statement #1: 'This teacher set out the course syllabus and the evaluation criteria clearly'; Statement #2: 'With this teacher, I learn'; Statement #3: 'This teacher motivates me to make an effort and to learn by myself'; Statement #4: 'The course material that the teacher provides me with helps'; Statement #5: 'The evaluation procedure allows me to demonstrate my knowledge'; and Statement #6: 'This teacher helped me overcome my doubts when I consulted him/her'.
Source: Author's computation using data from teaching evaluations at the University of Girona, 2018/19, 2019/20.

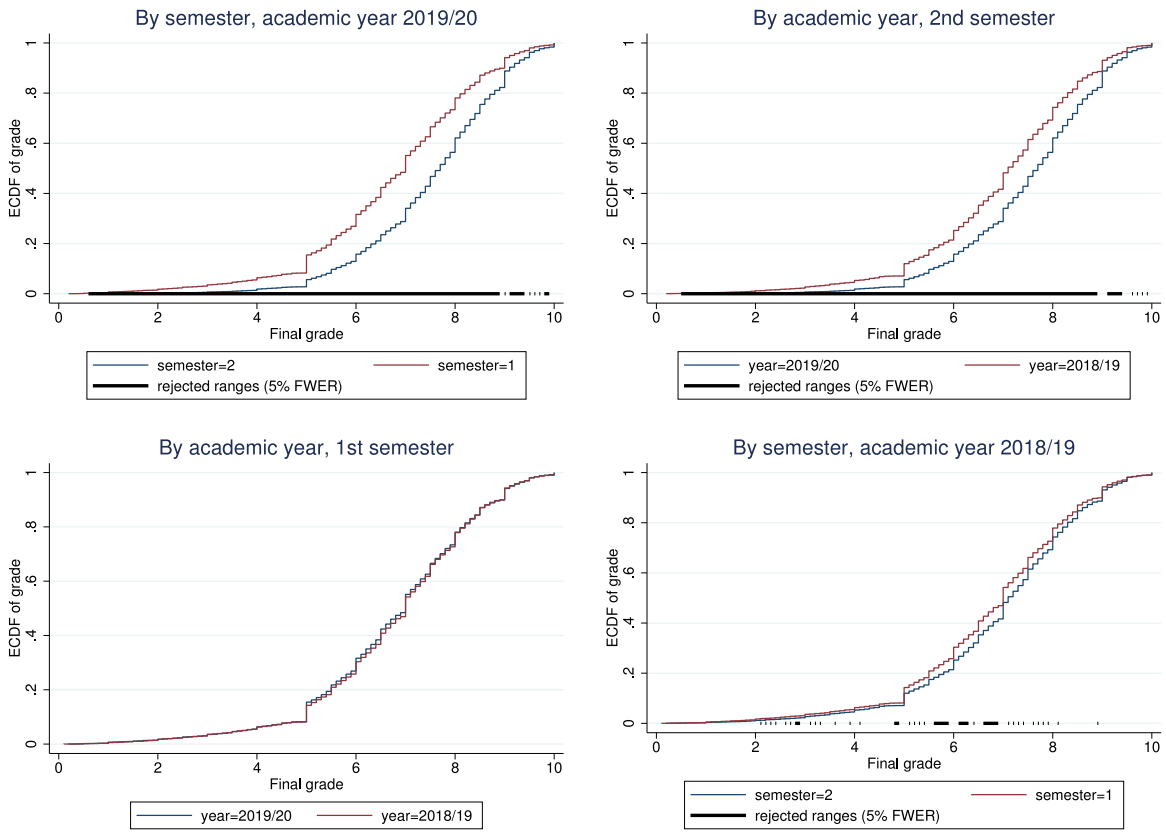

Fig. A.2. Empirical cumulative distribution functions (CDF) from final grades, academic years 2018/19 and 2019/20. Note: Stata's command *distcomp* by Kaplan (2019) was used for this analysis. FWER refers to family-wise error rate.
Source: Author's computation using data from teaching evaluations at the University of Girona, 2018/19 and 2019/20.

Table A.1
Balancing checks.
Source: Author's computation using data from teaching evaluations at the University of Girona, 2018/19 and 2019/20.

	Female student (1)	Student age (2)	Course repeater (3)	Student final grade (4)	Arts and Humanities (5)	Sciences (6)	Life Sciences (7)	Social Sciences (8)	Engineering and Architecture (9)
Online semester	0.0192	0.0236	−0.0063***	0.4855***	−0.0057***	−0.0021	0.0031*	0.0043*	0.0003
	(0.0135)	(0.1310)	(0.0023)	(0.0380)	(0.0017)	(0.0024)	(0.0016)	(0.0025)	(0.0026)
Academic year 2019/20	0.0178*	0.0274	0.0062***	0.0279	0.0022**	0.0006	0.0005	−0.0033*	−0.0000
	(0.0094)	(0.0783)	(0.0018)	(0.0262)	(0.0010)	(0.0014)	(0.0010)	(0.0018)	(0.0017)
2nd semester	0.0100	0.6929***	−0.0011	0.1386***	0.0125***	0.0300***	−0.0112***	−0.0108***	−0.0204***
	(0.0100)	(0.0970)	(0.0016)	(0.0302)	(0.0015)	(0.0021)	(0.0013)	(0.0021)	(0.0023)
Observations	78 070	78 070	78 070	77 141	77 612	77 612	77 612	77 612	77 612

Note: The dependent variable of each regression is detailed in the column header. Each regression includes teacher fixed effects. Standard errors are clustered at the student level. Significance level: ***p < 0.01, **p < 0.05, *p < 0.1.

Table A.2
Difference-in-differences results for teaching evaluations at the University of Girona by lecturer's gender — Robustness checks.
Source: Author's computation using the whole universe of teaching evaluations at the University of Girona, 2015/16 to 2019/20.

Panel A: From academic year 2015/16		
	Male (1)	Female (2)
Online semester	0.0021	−0.0404*
	(0.0202)	(0.0230)
Academic year 2016/17	−0.0060	−0.0061
	(0.0128)	(0.0136)
Academic year 2017/18	0.0231*	−0.0104
	(0.0137)	(0.0142)
Academic year 2018/19	−0.0312**	0.0047
	(0.0149)	(0.0152)
Academic year 2019/20	−0.0161	0.0444**
	(0.0183)	(0.0203)
2nd semester	−0.0644***	−0.0116
	(0.0102)	(0.0112)
Observations	111 568	88 696
Panel B: Placebo online semester		
Placebo online semester	0.0173	0.0455*
	(0.0247)	(0.0275)
Academic year 2018/19	−0.0579***	−0.0123
	(0.0166)	(0.0192)
2nd semester	−0.0606***	−0.0282
	(0.0196)	(0.0212)
Observations	44 378	35 801
Panel C: Adding fixed effects by degree		
Online semester	−0.0069	−0.0636**
	(0.0256)	(0.0292)
Academic year 2019/20	0.0201	0.0528***
	(0.0168)	(0.0202)
2nd semester	−0.0511**	0.0409*
	(0.0203)	(0.0226)
Observations	42 235	34 031

Note: Controls include student age, its square, student gender, course repeater and field of study. All regressions include lecturer fixed effects. Standard errors are clustered at the student level. Panel A includes data from academic year 2015/16 onwards. Panel B simulates a scenario whereby the online semester occurred in the second semester of 2018/19. Panel C adds fixed effects by degree and excludes the control for field of study. Significance level: ***p < 0.01, **p < 0.05, *p < 0.1.

A.2. Tables

See Tables A.1 and A.2.

References

Adams-Prassl, A., Boneva, T., Golin, M., & Rauh, C. (2020). Inequality in the impact of the coronavirus shock: Evidence from real time surveys. *Journal of Public Economics*, *189*, Article 104245.

Alon, T., Doepke, M., Olmstead-Rumsey, J., & Tertilt, M. (2020a). The impact of COVID-19 on gender equality. *Covid Economics: Vetted and Real-Time Papers*, *4*, 62–85.

Alon, T., Doepke, M., Olmstead-Rumsey, J., & Tertilt, M. (2020b). *This time it's different: The role of women's employment in a pandemic recession*: *CEPR Discussion paper no. 15149*.

Aucejo, E. M., French, J., Ugalde Araya, M. P., & Zafar, B. (2020). The impact of COVID-19 on student experiences and expectations: Evidence from a survey. *Journal of Public Economics*, *191*, Article 104271.

Aucejo, E. M., French, J. F., & Zafar, B. (2021). *Estimating students' valuation for college experiences*: *Working Paper 28511*, National Bureau of Economic Research.

Bagues, M. F., & Esteve-Volart, B. (2010). Can gender parity break the glass ceiling? Evidence from a repeated randomized experiment. *Review of Economic Studies*, *77*(4), 1301–1328.

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

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

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

Boring, A., Ottoboni, K., & Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, 1–11.

Boring, A., & Philippe, A. (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*, Article 104323.

Browning, M. H. E. M., Larson, L. R., Sharaievska, I., Rigolon, A., McAnirlin, O., Mullenbach, L., Cloutier, S., Vu, T. M., Thomsen, J., Reigner, N., Metcalf, E. C., D'Antonio, A., Helbich, M., Bratman, G. N., & Alvarez, H. O. (2021). Psychological impacts from COVID-19 among university students: Risk factors across seven states in the United States. *PLoS One*, *16*(1), 1–27.

Carrell, S., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy*, *118*, 409–432.

Deryugina, T., Shurchkov, O., & Stearns, J. (2021). COVID-19 disruptions disproportionately affect female academics. *AEA Papers and Proceedings*, *111*, 164–168.

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

Farré, L., Fawaz, Y., González, L., & Graves, J. (2022). Gender inequality in paid and unpaid work during Covid-19 times. *Review of Income and Wealth*, in press.

Fauville, G., Luo, M., Queiroz, A. C. M., Bailenson, J. N., & Hancock, J. (2021). *Nonverbal mechanisms predict Zoom fatigue and explain why women experience higher levels than men*: *Discussion paper*, SSRN.

Henning, M. A., Krägeloh, C. U., Dryer, R., Moir, F., Billington, D. R., & Hill, A. G. (Eds.), (2018). *Wellbeing in higher education: Cultivating a healthy lifestyle among faculty and students*. London: Routledge.

Hoffmann, F., & Oreopoulos, P. (2009). Professor qualities and student achievement. *The Review of Economics and Statistics*, *91*(1), 83–92.

Jaeger, D. A., Arellano-Bover, J., Karbownik, K., Martínez Matute, M., Nunley, J. M., Seals, R. A., Jr., Almunia, M., Alston, M., Becker, S. O., Beneito, P., Böheim, R., Boscá, J. E., Brown, J. H., Chang, S., Cobb-Clark, D. A., Danagoulian, S., Donnally, S., Eckrote-Nordland, M., Farré, L., ... Zhu, M. (2021). *The global COVID-19 student survey: First wave results*: *IZA Discussion Papers 14419*, Institute of Labor Economics (IZA).

Kaplan, D. M. (2019). Distcomp: Comparing distributions. *Stata Journal*, *19*(4), 832–848.

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

McElroy, M. B. (2016). Report: Committee on the status of women in the economics profession (CSWEP). *American Economic Review*, *106*(5), 750–773.

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

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.

Rodríguez-Planas, N. (2022a). COVID-19, college academic performance, and the flexible grading policy: A longitudinal analysis. *Journal of Public Economics*, *207*, Article 104606.

Rodríguez-Planas, N. (2022b). Hitting where it hurts most: COVID-19 and low-income urban college students. *Economics of Education Review*, *87*, Article 102233.

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

Sarsons, H., Gërxhani, K., Reuben, E., & Schram, A. (2021). Gender differences in recognition for group work. *Journal of Political Economy*, *129*(1), 101–147.

Sinclair, L., & Kunda, Z. (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.

Stark, P., & Freishtat, R. (2014). An evaluation of course evaluations. *ScienceOpen Research*, *9*, 1–7.

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

Zamarro, G., Camp, A., Fuchsman, D., & McGee, J. (2022). *Understanding how Covid-19 has changed teachers' chances of remaining in the classroom*: *Working Paper 22–01*, Sinquefield Center for Applied Economic Research.

Zamarro, G., & Prados, M. (2021). Gender differences in couples' division of childcare, work and mental health during COVID-19. *Review of Economics of the Household*, *19*, 11–40.