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

DISCUSSION PAPER SERIES

IZA DP No. 14387

Gender Differences in Student Evaluations of Teaching: Identification and Consequences

Edmund Cannon
Giam Pietro Cipriani

MAY 2021

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14387

Gender Differences in Student Evaluations of Teaching: Identification and Consequences

Edmund Cannon
University of Bristol

Giam Pietro Cipriani
University of Verona and IZA

MAY 2021

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 Differences in Student Evaluations of Teaching: Identification and Consequences

Student Evaluations of Teaching (SETs) have been suggested as one possible cause for low representation of women among academic economists. While econometric analyses using control variables certainly report that SETs can be influenced by the gender of both teacher and student, such studies may still be biased if there is non-random allocation of teachers to teaching. Even if causal estimates of gender effects are unbiased, the inference that SETs contribute to gender discrimination is hazardous, since hiring or promotion committees would not have access to these controls when evaluating SETs. We use data from an Italian university to quantify the effect of controls on gender effects and conclude that there is insufficient evidence to blame SETs for a gender imbalance in Economics.

JEL Classification: A22, I21, J16

Keywords: students' evaluation of teaching, gender bias, matching, teaching allocation, hiring and promotion

Corresponding author:
Giam Pietro Cipriani
Department of Economics
University of Verona
Polo Santa Marta
Via Cantarane, 24
37129 Verona
Italy

E-mail: giampietro.cipriani@univr.it

1. Introduction

There is increasing concern at the slow progress in increasing the proportion of female academic staff in Economics (Bayer and Rouse, 2016).[1] Student Evaluations of Teaching (SETs) have been identified as a potential contributor to this problem because they are widely used to measure teaching within Economics departments and have consequences for hiring and promotion decisions (Becker and Watts, 1999; Becker et al., 2012).[2] The analyses of McPherson et al. (2009), Wagner et al. (2016), Boring (2017) and Mengel et al. (2019) all find evidence of gender effects in SETs and infer that this puts female teaching staff at a disadvantage. In this paper we make two contributions to this literature.

First, we note that a hiring or promotion decision is likely to be based on a simple average SET score for the member of staff, but analyses of the determinants of SETs are based on the individual students' SET responses plus additional control variables and sophisticated identification strategies. Estimated gender effects will differ depending on the controls and methods used and it is often difficult to infer the magnitude of the gender bias potentially facing hiring and promotion committees from published papers on SETs. In this paper we use a straightforward analysis to lay out the relationship between the two gender effects and use an empirical example to show that the effects may be different. We also discuss whether any gender effect is sufficiently large for it to be a major contributor to gender biases in the Economics profession.

Second, several recently-published papers have used multi-section studies to obtain better estimates of gender biases in SETs and multi-section studies have been held up a superior to alternative identification strategies. However, multi-section studies often fail to discuss whether there is random allocation of teachers to teaching. Without such random allocation multi-section studies might still be biased; if the allocation of

[1] Any discussion of gender raises the issue of transgender individuals, but it is rare for existing publications to discuss this, probably due to a lack of data on this personal information. In our empirical analysis, teachers are treated as female if they have a female first name.

[2] The possibility of gender bias in promotion and hiring in economics might have further important consequences given that male and female academic staff in Economics appear to have systematically different opinions about economic policy (May et al, 2014), but such issues are outside the scope of this paper.

teachers were completely random then it is not clear why controls are needed anyway. It seems plausible that teaching allocation is non-random, raising the question of how teaching might be allocated. Furthermore, if the teaching allocation is non-random it is possible that knowledge of this will be used to offset gender biases in SETs when making hiring and promotion decisions (if the teaching allocation is completely random, then it is to be hoped that this does not suggest that the hiring and promotion decisions are completely random too).

The rest of our paper is structured as follows. In section 2 we provide a short review of the literature to discuss how to estimate gender effects in SETs. In section 3 we show the relationship between the average SETs teaching staff get on a unit and the individual students' SET responses. Section 4 discusses the issue of teaching allocation and shows that an optimal teaching allocation could result in gender effects even in a multi-section study. Section 5 describes our data and section 6 provides our empirical example. Section 7 concludes.

2. Literature review of gender effects

The literature on SETs is vast (Al-Issa and Sulieman 2007 report that there are thousands of papers) and it is well known that SETs are highly problematic (e.g. Spooren et al., 2013). Since our focus is on gender we do not attempt to discuss all of the issues here. However, we cannot entirely avoid the issue of whether SETs are good indicators of the resulting learning by students (usually measured by assessment outcomes), since we need to be confident that gender effects in SETs are not due to the quality of the teaching. In fact the most recent meta-analysis of the relationship between SETs and student learning suggests that the correlation between SET scores and student learning is negligible (Uttl et al., 2017). This result is also found consistently in papers confined to Economics, namely Braga et al. (2014), McPherson et al. (2009), Boring (2017), Mengel et al. (2019) and Linask and Monks (2018) all of which suggest that the gender of the teacher has no implications for student learning. The few studies that have attempted to find a correlation between SET scores and longer-term learning (measured by grades on subsequent units) also find no relationship (Weinberg et al., 2009; Carroll and West, 2010). Further problems with SETs may arise because students' responses change over time (Clayson, 2018; Vanacore & Pellegrino, 2019) or depend on whether the teacher is good looking or likable (Feistauer and Richter, 2018).

In this paper we confine our attention to the possibility of gender bias. For most of the analysis we shall assume that students respond to questions with a Likert scale and that the numerical scores are averaged to calculate what we shall call the "unit SET". It is well known that averaging ordinal responses in this way is inappropriate but the method is used by many universities.

The meta-analysis of Feldman (1993) found a statistically significant bias against women but the magnitude of the effect was negligible. In a large study based on the evaluations of 741 lecturers in 1995-96 from 21 different higher educational institutions in the USA, Centra and Gaubatz (2000) reported that female teaching staff performed slightly better than male staff overall, although the effect differed across subject areas. This paper also found that gender effects depend not just on the gender of the teacher but also the gender of the student: male students' comparative evaluation of male and female teachers differs from that of female students. A consequence is that the overall SET score received by a teacher depends not only on the individual gender biases of the students doing the evaluations, but also potentially on the gender composition of the students doing the evaluation (which may itself be correlated with the gender of the teacher).

The simplest way to test whether there is gender bias in SETs would be to compare unit SETs of male and female teachers controlling for student gender, but this would fail to control for other causal factors. This approach can still be informative: for example Basow and Montgomery (2005) use it to compare student evaluations of teaching with staff's self evaluation. However, most papers attempt to include control variables to get unbiased estimates of the determinants of SETs. The potential problems can loosely be characterized as falling into three groups:

i. gender may be correlated with other characteristics which influence student responses teaching (including response rates);
ii. teaching staff need not be randomly allocated to students;
iii. staff need not be randomly allocated to different forms of teaching.

Most papers attempt to address the first issue by controlling for observed teacher and student characteristics, such as age, experience or ethnicity of teacher; age of student; or students' prior exam grades (some or all of these covariates are used by Boring, 2017; and Mengel et al., 2019). The number of control variables available is usually quite small, but it is rare for papers to rely entirely upon control variables to estimate the

determinants of SETs. The only recent paper that we have found analysing SETs from Economics students using just control variables is Fenn (2015) and that study is more about the relationship between responses to different questions on the SET.

One approach to overcome a paucity of control variables in this context is to use a multi-section study and Abrami et al. (1990) characterize these as the strongest validation method for assessing whether SETs measure teaching effectiveness. Uttl et al. (2017, p.23) characterize a multi-section study follows: "a course has many equivalent sections following the same outline and having the same assessments, students are randomly assigned to sections, each section is taught by a different instructor, all instructors are evaluated using SETs at the same time and before a final exam, and student learning is assessed using the same final exam." Furthermore, since the timetabling systems used to create sections can be described, it is usually possible to argue that there is random allocation of students to teachers and so this method also solves the second of the three problems we have just identified.

Recent papers which use a multi-sectional analysis of SETs within Economics include studies from French and Dutch universities (Boring, 2017; Mengel et al., 2019). In both instances, the SETs being analyzed are for teaching sessions on a large unit with many students:[3] for example, two hundred students might receive lectures from a professor but then be divided into ten groups of twenty students each to be taught in seminars. The internal validity of the studies is potentially strong because SETs are observed for different staff teaching the same material on the same unit, so there is less need to find alternative controls for confounding factors which might influence SETs. In the case of Boring (2017), it is also possible to see how the same student evaluates different members of staff, effectively controlling for student heterogeneity as well. However, multi-section studies are not a panacea for analysing SETs since it is not always possible to implement them. For example, in both of the papers just mentioned the small-group seminar teaching can be analyzed using this method, but the large-group lectures cannot (because the lectures are only delivered once). Large-group lectures could only be

[3] To avoid confusion: by "unit" we mean a course of teaching with its own assessment; a degree consists of units. Most of the degrees analyzed in this paper are three-year undergraduate degrees with six units per year.

analyzed by multi-section studies where lectures of the same material are delivered multiple times by different members of staff.

An alternative way to try to control for the effects of unit content on SETs is to analyze SETs where two or more staff are co-teaching on the same unit. Wagner et al. (2016, p.81) describe co-teaching on Social Science units, where the co-teaching can consist of staff sharing out teaching sessions or teaching jointly in the same session. However, allocation of staff to units appears to be non-random and it is unclear how the different topics and material within a unit distributed between staff (the mechanism is described as "informal"). An analogous study of SETs in Chemical Engineering by Hempel et al. (2019) appears to address the second issue because all sessions were co-taught with both teachers simultaneously in the room and using (or trying to use) the same lecturing style. The Likert-scale responses for the SET showed lower scores for the female teacher than the male teacher but the verbal response answers showed no gender bias. This suggests that co-teaching may reduce the problem of gender-specific comments found by Schmidt (2015) in his analysis of verbal responses on RateMyProfessor.com.

A different approach again is taken by Bavishi et al. (2010), who ask college preparatory students to evaluate professors on the basis of hypothetical CVs. The advantage of this is that there are no confounding factors at all; conversely the analysis is more informative about gender bias in general than gender bias in SETs. This paper finds no statistically significant effect for gender overall, but does find lower approval for female tutors from ethnic minorities.

With the exception of Bavishi et al. (2010), all of these papers need random allocation of teachers to different types of teaching (the third issue that we identified above). Our reading of published multi-sectional studies is that this issue is not really addressed. We think it uncontroversial to suggest that different tutors have strengths in different areas or types of teaching and the authors' experience of leadership within two different economics departments has convinced them that it is common for academic colleagues to have comparative advantage. This results in non-random allocation of staff to different types of teaching: the question then is whether (or under what circumstances) particular teaching allocations could result in gender effects even in multi-section studies. We discuss this in detail in section 4.

3. Gender effects and hiring or promotion decisions

In this section we set out a conceptual framework to describe how differences in the unit SET scores of teaching staff (which might plausibly affect hiring and promotion) arise from gender effects in the SET choices of individual students (which are the typical subject of analyses of SETs). It is obvious that the two effects will usually be different: the purpose of this section is to decompose the difference to facilitate comparison of the two: we shall provide a numerical example in section 5.

There is no systematic evidence on how hiring and promotion committees use information on SETs (Linse, 2017). Anecdotal evidence (e.g. Mengel et al, 2017, p.542) suggests that it is common for the unit SET score $\bar{y}_t$ to be used for hiring and promotion decisions, where

$$\bar{y}_t \equiv n^{-1} \sum y_{t,s} \tag{1}$$

and $y_{t.s}$ is the SET score given by student s on unit t and we refer to $\bar{y}_t$ as the "unit SET". We have already noted that averaging responses to Likert-scale questions like this is strictly invalid, but we shall assume that this is the relevant measure for our analysis.

A naïve analysis of gender effects would simply compare the mean unit SETs for male and female teachers. To relate this method to our later discussion we note that this is the same as a regression analysis of the form

$$\bar{y}_t = \alpha'_0 + \alpha'_1 f_t + \eta'_t \tag{2}$$

where f_t is an indicator dummy variable taking the value one if the unit lecturer is female and zero otherwise and an estimate of α'_1 quantifies the difference between the two means and will be negative if female teachers get lower unit SETs. (In principle one could be interested in other parts of the distribution, but in this paper we shall concentrate on the mean). Causal analyses of gender would typically use the individual student SETs, modelled as

$$y_{t.s} = \alpha_0 + \alpha_1 f_t + \eta_{t.s} \tag{3}$$

and f_t is an indicator dummy variable taking the value one if the teacher is female and zero otherwise.The relationship between α'_1 and α_1 depends on such issues as whether female staff tend to get bigger or smaller units on average and whether size of unit influences the students' responses. We discuss this issue in our empirical example in

section 5, where we find the difference to be small. We have seen no discussion of this issue in published papers.

Of course, a simple adjustment of α_1 would fail to control for other reasons why male and female staff received different SETs, but it is unlikely that a promotions committee would have a full set of conditioning information and even less likely that a hiring committee would be able to control for differences between the students and institutions from which it was receiving applications. If the only information available to a committee to help it interpret the SET is the candidate's gender, then α_1 (or α_1') is the relevant piece of information since

$$\alpha_1 = \mathbb{E}[y_{t.s}|f_t = 1(female)] - \mathbb{E}[y_{t.s}|f_t = 0(male)]. \quad (4)$$

Empirical analyses to estimate gender effects typically control for other factors using a model such as

$$y_{t.s} = \beta_0 + \beta_1 f_t + \beta_2 f_s + \beta_3 f_t f_s + \boldsymbol{\mu}_1 \boldsymbol{X}_t + \boldsymbol{\mu}_2 \boldsymbol{X}_s + \varepsilon_{t.s} \quad (5)$$

where f_t is as defined above; f_s is an analogous dummy variable for the gender of the student; and $\boldsymbol{X}_t$ and $\boldsymbol{X}_s$ are vectors of control variables for the teacher and student respectively. If one can obtain unbiased parameter estimates, the interpretation of the gender-specific parameters is as follows:

Hypothesis	Test
No gender differences	$\beta_1 = \beta_2 = \beta_3 = 0$
Male students evaluate male and female teachers the same (vs male students evaluate female teachers lower)	$\beta_1 = 0$ vs $\beta_1 < 0$
Female students evaluate male and female teachers the same (vs female students evaluate female teachers lower)	$\beta_1 + \beta_3 = 0$ vs $\beta_1 + \beta_3 < 0$
Male teachers are evaluated the same by male and female students (vs male teachers are evaluated more highly by female students)	$\beta_2 = 0$ vs $\beta_2 > 0$
Female teachers are evaluated the same by male and female students (vs female teachers are evaluated more highly by female students)	$\beta_2 + \beta_3 = 0$ vs $\beta_2 + \beta_3 > 0$
Differences in evaluation of male and female teachers do not depend on student gender	$\beta_3 = 0$

Notice that if the model were specified not in terms of female indicator variables but male indicator variables (taking the value one when respectively the tutor or student is male and zero otherwise) then the model would be re-written as

(6) $$y_{t.s} = \beta_0' + \beta_1' m_t + \beta_2' m_s + \beta_3' m_t m_s + \boldsymbol{\mu}_1 \boldsymbol{X}_t + \boldsymbol{\mu}_2 \boldsymbol{X}_s + \varepsilon_{t.s}$$

where

(7) $$\beta_1' = -(\beta_1 + \beta_3),\ \beta_2' = -(\beta_2 + \beta_3),\ \beta_3' = \beta_3.$$

From equation (3), the difference in the average expected SET for male and female teachers is

(8) $$\mathbb{E}[y_{t.s}|f_t = 1] - \mathbb{E}[y_{t.s}|f_t = 0] = \beta_1 - \beta_2 \mathbb{E}[f_s|f_t = 0] + (\beta_2 + \beta_3)\mathbb{E}[f_s|f_t = 1]$$
$$+\boldsymbol{\mu}_1 \mathbb{E}[\boldsymbol{X}_t|f_t = 1] - \mathbb{E}[\boldsymbol{X}_t|f_t = 0] + \boldsymbol{\mu}_2 \mathbb{E}[\boldsymbol{X}_s|f_t = 1] - \mathbb{E}[\boldsymbol{X}_s|f_t = 0]$$

The problem with these specifications is that the number of control variables $\boldsymbol{X}_t$ and $\boldsymbol{X}_s$ might be quite limited and unlikely to explain much variation in student responses. Variables such as age of student have little variation; age of teacher varies more, but then there is an issue of the appropriate functional form. Where students come from different educational systems (perhaps in different countries) it is difficult to control for students' prior exam grades (some or all of these covariates are used by Wagner et al., 2016; Boring, 2017; and Mengel et al., 2019). Pre-determined exam grades may be statistically endogenous due to correlation with unobserved characteristics such as unobserved student propensity to work. To reduce these problems one can use either teacher fixed effects or student fixed effects,[4] respectively

(9) $$y_{t.s} = \gamma_0 + \gamma_1 m_t m_s + \gamma_2 f_t f_s + \sigma_s + \boldsymbol{\lambda_1} \boldsymbol{X}_t + \phi_{t.s}$$

and

(10) $$y_{t.s} = \delta_0 + \delta_1 m_t m_s + \delta_2 f_t f_s + \tau_t + \boldsymbol{\lambda}_1 \boldsymbol{X}_s + \psi_{t.s}$$

where the correspondence between the gamma and delta parameters and the betas is as follows:

γ_1	Difference in male student evaluation of male v. female tutor.	$-\beta_1$
γ_2	Difference in female student evaluation of male v. female tutor.	$\beta_1 + \beta_3$
δ_1	Difference in male tutor evaluation by male v. female student.	$-\beta_2$
δ_2	Difference in female tutor evaluation by a male v. female student.	$\beta_2 + \beta_3$

[4] By teacher fixed effects we mean a complete set of indicator variables, where for fixed effect t, the variable takes the value one for tutor t and zero otherwise. Student fixed effects are defined analogously. Linask and Monks (2018, p.326) report that very few papers use fixed effects as controls.

Linask and Monks (2018) consider model (9) but omit the interactive female term $f_t f_s$. Boring (2017) considers both specifications within a multi-section analysis.

Unfortunately, it is impossible to include both sets of fixed effects in the same regression. To see why, consider the interpretation of the parameter δ_1 which is associated with the indicator variable $m_t m_s$, which takes the value one when both teacher and student are male and is interpreted as a partial effect, holding everything else constant. For this variable to take the value zero either the teacher would have to be different (which would be consistent with student fixed effects but not teacher fixed effects) or the student would have to be different (which would be consistent with teacher fixed effects but not student fixed effects).[5] With these models it is possible to estimate β_1 or β_2 but not both and hence one can only calculate the overall gender effect if one has complete knowledge of all of the estimated fixed effects.

This is just a specific example of the fact that it is difficult to infer much about α_1' or α_1 from knowledge of the beta, gamma or delta parameters. A causal analysis of gender effects typically discusses whether the parameter estimates of the betas (or gammas or deltas) are large or statistically significant whereas the relevant information for the policy question is whether the alphas are large relative the overall variation in unit SETs, usually measured by the standard deviation. Published research does not always discuss this fully. For example, Boring (2017) finds statistically significant gender effects but does not report the relevant standard deviation of the unit SETs.

Mengel et al. (2019) do report the standard deviation of unit SETs and find that the effect of male students' SETs would result in female teachers getting 20.7% of a standard deviation less than male teachers. The actual differences in unit SET scores are not reported: however, for a hypothetical unit with equal numbers of male and female students, female teachers would only get 14.2% of a standard deviation of a SET score less than male teachers. Wagner et al. (2016) also find statistically significant gender effects but the overall difference in unit SETs of male and female teachers is only 6.5% of a standard deviation. There is no discussion of whether these effects are sufficiently large to influence hiring or promotion committees or to be interesting determinants of

[5] An alternative way of describing this problem is that attempting to include both teacher and student fixed effects alongside $m_t m_s$ or $f_t f_s$ would result in multicollinearity.

the low proportion of female staff in the Economic profession. Linask and Monks (2018, p.331) find an effect size for $m_t m_s$ (i.e. the parameter γ_1) which is a half of the standard deviation of unit SETs and we accept that this is large: unfortunately it is impossible to interpret since they omit the corresponding explanatory variable $f_t f_s$.

4. Teaching allocation and multi-section studies

We now turn to the effect of the allocation of teachers to teaching. Prima facie this needs to be random for analyses of SETs to estimate gender effects, but the issue is rarely discussed in published papers. We wish to know under what circumstances a teaching allocation could lead to biased estimates and precisely how this might affect multi-section studies.

We shall assume that there is some underlying student satisfaction which is measured by SETs (although it may have little to do with learning or anything else). Throughout this section we shall define a "better" teacher as one who tends to obtain higher SET scores (i.e. the objective of the department is maximising measured student satisfaction, not student learning). A key point in our argument is that teachers are unlikely to be allocated randomly to teaching (Ragan and Walia, 2010) and that there may be comparative advantage in increasing student satisfaction in different types of teaching. We shall use two simple examples to show that the combination of these assumptions can lead to prima facie counter-intuitive results for aggregate SETs. For ease of exposition we ignore the fact that male and female students may evaluate differently.

In the first example, we create a hypothetical example where female teacher evaluations could be lower than male teacher evaluations (consistent with the data), despite female teachers being "better" than male teachers in the sense just defined. We assume that the person allocating the teaching aims purely to maximize measured student satisfaction and does not vindictively allocate teaching to lower female teachers' scores.

Suppose that there are just two types of teachers (1 and 2); two types of unit (A and B); in this example we also assume that there are equal quantities of teacher and unit types. Furthermore, we assume that there is a correlation between teacher type and gender, so that type 1 teachers are predominantly male and type 2 teachers are predominantly female (the possibility of teaching type being correlated with gender is consistent with the evidence of Boring, 2017, who finds that female teachers do get higher scores on

certain aspects of teaching despite getting lower scores overall). Finally, we assume that students do not discriminate on gender grounds and complete SETs purely on the basis of the perceived ability of a given type of teacher for a given type of teaching.

To complete this hypothetical example, we model expected SET scores as follows:

	Gender of teachers	Unit/teaching type A	Unit/teaching type B
Teacher type 1	predominantly male	3.6	2.8
Teacher type 2	predominantly female	3.8	3.4

In this example, teacher type 2 is unambiguously better than teacher type 1 in terms of maximising student satisfaction, because they would get better scores whatever they taught. If teachers were allocated randomly, then the expected score for type 1 teachers would be 3.2, the expected score for type 2 teachers would be 3.6 and the expected average score for the department would be 3.4. A disadvantage of this allocation is that a quarter of the teaching would be given the relatively low SET of 2.8.

Suppose, however, that teachers are not randomly allocated: the optimal allocation of teaching is based on comparative advantage, putting all type 1 teachers on A units and all type 2 teachers on B units: the average score would then be 3.5 and no teaching would get a score lower than 3.4. The paradoxical result would be that type 2 teachers would now get 3.4, less than the 3.6 received by their inferior type 1 colleagues. Since we have assumed that type 2 teachers are predominantly female, the average SET score of female teachers would be less than the average SET score of male teachers, despite the fact that the students think they are better teachers and evaluate without discrimination.

The first conclusion of this argument is that the SETs received by individual teaching staff may not even relate closely to their ability to raise student satisfaction (let alone other desiderata such as student learning). The corollary is that identifying student gender bias in SETs from gender effects is also highly problematic.

Notice that the model frameworks presented in section 3 would be inappropriate to model this scenario: all of those models assume that if an observable characteristic explained SETs then that characteristic would have a homogeneous treatment effect on all sorts of teaching (whereas comparative advantage suggests that there are heterogeneous treatment effects). Fixed-effect specifications suffer from this problem

just as much as the models with controls: the estimated teacher fixed effect shows how much better a member of staff is at all forms of teaching (the effect is fixed across different units and subjects).

As discussed in section 3, an additional way to identify gender effects is to use a multi-section study and we now turn to the effect of non-random teaching allocation on this

The first hypothetical example is less helpful for our understanding of the effect of the teaching allocation on multi-section studies. In the example each type of teaching was only allocated one type of tutor and this allowed for no possibility of intra-unit variation, precisely the variation that multi-section studies measure. To approach this problem, we need to think why multi-section studies might find intra-unit variation in SETs within the context of an optimal teaching allocation.

We turn to a second example to show that optimal teaching allocation could lead to a variety of conclusions from a multi-section study. Suppose we consider a different matrix of student satisfaction, as follows:

	Gender of tutor	Unit type A	Unit type B
Teacher type 1	predominantly male	3.4	3.8
Teacher type 2	predominantly female	3.8	3.4

In this example, neither teacher type dominates the other, but there is still comparative advantage. Unlike the first example we do not assume anything about the shares of teacher or teaching types.

If the shares of the two types of teaching match the shares of the appropriate teacher types then an optimal allocation ensures that type 2 teachers do type A teaching, type 1 teachers do type B teaching, all SETs are 3.8 and there will be no gender effect. However, it is quite possible that the types of teachers do not match the types of teaching. If there are too few type 1 teachers then some type 2 teachers must do type B teaching and a multi-section study will conclude that type 2 teachers get lower SETs; since type 2 teachers are predominantly female, the SET scores of female tutors will be lower on

average in type B units and multi-section studies will also find a gender effect where female tutors get lower scores.[6]

These examples highlight the fact that while non-random teaching allocation might result in apparent gender effects in the data, precisely what those effects will be depend upon both the allocation and the method of analysis. In terms of the allocation, the sort of intra-unit variation which would be measured by multi-section studies depends partly upon the constraints faced in the allocation, which may depend upon the size of department of the range of subject specialisms among the staff.

5. Data and the Italian university system

We collected data for all student evaluations for an Italian university. The only other detailed study of the determinants of SETs in an Italian university appears to be Braga et al. (2014) based on data from Bocconi, which is an elite university and possibly not entirely representative.

Students were not allowed to attempt the summative assessment (typically an exam) until they had completed the evaluation. The SETs were administered anonymously, and only summary statistics were made available to the lecturers. However, enough information was available to match student evaluations to other pieces of information such as student gender. Table 1 summarizes our data set: we have information on 727 units but in many cases the number of student responses is low: this is largely driven by the fact that some optional units have very few students enrolled. We shall usually exclude the smallest units by confining our analysis to units with more than ten SETs (where the choice of ten is arbitrary but makes no substantive difference to the results). We discuss the issue of Attenders and Non-attenders below.

When we break down the results by Faculty, we see that the majority of teachers are male, but the effect is much more pronounced in Business and Sport. Within the

[6] This assumes that there are "too many" type 2 tutors, who are predominantly female, which might seem unlikely given the low proportion of female in subjects such as Economics. But note that the excess of type 2 tutors is conditional on the split between type A and type B teaching: if male tutors predominate then most academic staff will be type 1 and they may bias the syllabus to type B teaching. This could be consistent with controversies within the economics profession about both the content and delivery of the subject (Bowles and Carlin, 2020) and the fact that women tend to have different beliefs to men about Economics (May et al., 2014).

Business Faculty a proportion of 29 per cent female teachers is not out of line with the corresponding number of 23.5 per cent reported in Economics departments in the USA (Bayer and Rouse, 2016). This is consistent with Economics having even more of a gender imbalance than Science subjects, although the numbers of staff here are too small to draw strong conclusions.

< Table 1 – Descriptive statistics – about here >

At the outset we need to clarify peculiarities of much of the Italian university system compared to that of other countries. In many systems educational provision consists of both large-group teaching sessions (or "lectures") and relatively small-group teaching sessions (or "seminars"). For the vast majority of the units and SETs that we are analysing, teaching consisted entirely or almost entirely of large-group teaching; where small-group teaching was present data on this were unavailable. Hence all of our SETs are for the large-group teaching component (although where the total number of students enrolled on a unit is small, the large-small distinction may be moot).

There are two other aspects of our data which are important. First, although students must complete the SET to take the final assessment, this does not guarantee a hundred per cent response rate in our data, since about one third of those enrolled will never attempt the exam and will never graduate. This raises the possibility that our data might have a selection effect by quality of student and low response rates might create systematic biases (Kherfi, 2011; Goos and Salomans, 2017; He and Freeman, 2020). However, the requirement that students complete the SET before they are allowed to sit the exam means that we have a complete set of responses from all students except those who are only minimally engaged with their course. University fees are very low in Italy and many students may enrol purely as a fallback option while trying to get a job: for a detailed discussion of selection effects in the Italian university context see Cipriani and Zago (2011). A second selection effect could arise from students choosing optional (elective) units on the basis of SETs or other information (Babad and Tayeb, 2003; Brown and Kosovich, 2015). This problem may be mitigated in the current study because many of the students whose SETs we are analysing are studying on degrees where a high proportion of units are compulsory (for example, this is true of the undergraduate Economics degrees) and choice of unit is effectively taken prior to matriculation (by

choosing either the degree of Economics Business or the degree of Economics and Accounting). There is no obvious way to address this issue given our data set.

The final peculiarity of the Italian system is that students chose to complete one of two different SETs; one was for students who had attended the lectures and one for those who had not (the latter evaluation allowed for the possibility that students had attended a small number of some of the teaching sessions). Lalla and Ferrari (2011) discuss problems with the response bias of the web survey to collect SET data and Liu (2012) finds evidence that the responses of distance learners demonstrate less of a gender effect. Wolbring and Treischl (2016) analyze the effect of attendance on SETs by conducting a SET twice during the same lecture course and seeing whether responses of students present on both occasions were systematically different from those students only present once. The analysis concluded that there was no effect on average but that there was some effect on course rankings. Unfortunately, the non-attendance analyzed by Wolbring and Treischl is different from ours: they are considering the possibility of students with lower (occasional) attendance whereas the formal definition of non-attenders in the Italian system is those who are effectively studying without using the lectures. Students in our data set were asked why they attended or did not attend and reasons for non-attendance are given: the three main reasons are attendance of other courses, work or "other" (details are reported in the Appendix). There is little more that we can glean from this but, as we have already noted, we have student evaluations of non-attendance for all those who wish to take the exam.

For those students who declared that they did attend lectures, we have responses to twenty questions, of which one is completely open for students' responses and we ignore (the vast majority of students left this blank). Most of the remaining questions allowed students to respond on a Likert scale of 1 (bad) to 4 (good), so the number of options on these questions is relatively small. Rivera and Tilcsik (2019) find that female tutors get better scores when there are six options rather than ten. A possible reason for this is that having fewer options benefits women because the nature of gender discrimination makes it hard to give a perfect "ten out of ten" to a woman; as the number of options is reduced the highest grade is less associated with being "perfect" and so students find it easier to award the top score to a woman. This suggests that having a Likert scale with only four points will tend to reduce the size of gender effects in our study.

Although we have responses to nineteen questions, in this paper we concentrate on results for overall student satisfaction, because student responses are highly correlated across different questions. In Cannon and Cipriani (2021) we show that this is partly due to "halo effects", apart from correlation of the underlying items. Here we assume that the overall-satisfaction question is an adequate summary of the SET as a whole.

Our data consist of SETs at a unit level, but a significant minority of units have more than one lecturer, in some cases a mixture of male and female teachers. To get a homogeneous data set and to enable us to relate the SETs to the lecturer's gender we confine our analysis to those units with just one lecturer. A further problem is that many university staff lecture on more than one unit, raising the question of whether we should concentrate on evaluations of an individual lecturer across all units on which he or she lectures or should concentrate on evaluation of a lecturer on each particular unit. Because one can imagine that a lecturer might have different skills on different units (someone teaching both macroeconomics and econometrics might be perceived as good at one subject but bad at the other), we treat each unit-lecturer combination as a separate observation. Similarly, in the very small number of instances where a unit is double lectured, we treat it as two separate units, even if the two groups have the same unit code.

Summary statistics are provided for the unit SETs in Table 2, which reports the mean for each gender and also the difference (the statistic in column 7 corresponds to α_1' in equation 2). Across the whole university female teachers get higher SETs, although the effect is small except in the sub-sample of Master students. However, there is much more variation when the sample is divided by Faculty and this suggests that our analysis should be at Faculty rather than university level. It is notable that female teachers get lower SETs in Business, but all of the within-Faculty gender differences are statistically insignificant due to the relatively small number of units being analyzed. We shall discuss this further in our regression analysis below.

< Table 2 - Summary statistics of Unit SET scores – about here >

6. Empirical example

We now use our data set to quantify the various effects discussed in section 3 applied to data at Faculty level (University-level analysis is reported in the Appendix). We compare

the differences in the mean SET between male and female teachers at both unit and student level; we then consider the effects of student gender; finally, we estimate fixed-effect regressions. As discussed in section 5, given that way that teaching is organized we are unable to conduct multi-section studies.

One of the decisions that must be taken when estimating models using student-level data is how to calculate the standard errors (conventional standard errors will be small due to the large sample sizes). This is important in our case because we have very few control variables: even if the omission of potential causes of student satisfaction does not result in omitted variables bias, it is still highly likely that $cov[\eta_{t.s}, \eta_{t.s'}] > 0$ for students $s \neq s'$ (if the tutor on a unit is perceived to be a good tutor by one student then they are likely to be perceived to be a good tutor by other students). Conversely we also know that there is significant student heterogeneity, so it is also possible that $cov[\eta_{t'.s}, \eta_{t.s}] > 0$ for units $t \neq t'$ if individual students systematically evaluate in the same way across units. However, the average number of SETs completed per student is just over three and over half of students only complete one or two SETs, so we think it likely that the intra-unit correlation is probably more significant and for this reason we choose to cluster the standard errors at unit level (where a lecturer teaches on more than one unit, we cluster these separately).

Our results from these simple regressions are reported in panels A and B of Table 3, which looks at the SETs from all 557 units for which more than ten students completed the evaluation, broken down by Faculty.

For most Faculties, the parameter estimates $\hat{\alpha}_1'$ and $\hat{\alpha}_1$ are typically very close to each other, consistent with female teachers not systematically teaching larger or smaller units than male teachers. Recall that this issue is relevant because hiring and promotion committees typically see unit SETs (where the difference is $\hat{\alpha}_1'$) while published analyses of SETs are typically conducted on individual students' responses (where the difference is $\hat{\alpha}_1$). In education the student-level analysis suggests very slightly higher SETs for male teachers whereas the unit SETs suggest higher SETs for female teachers, but both are statistically insignificant.

With the exception of the regression for the sub-sample of Masters students, the coefficient on the female dummy is statistically insignificant at conventional levels

regardless of whether we estimate the gender effect by unit or by SET. This is primarily because the effect size is small (and we have used conservative standard errors).

The gender effects in panels A and B are based on the effect of gender conditioning on no other variables at all. Corresponding regressions conditioning on gender of student and other controls are reported in panel C, which also report estimates of $\beta_1 + \beta_3$ and $\beta_2 + \beta_3$. Estimates of $\beta_1 + \beta_3$ are typically very small and statistically insignificant. There is more consistent evidence that $\beta_2 + \beta_3 > 0$ suggesting that female teachers are evaluated more highly by female students than by male students, although we have not corrected the significance of the tests for multiple testing. Finally consider a joint test that $H_0: (\beta_1 = 0) \cap (\beta_2 = 0) \cap (\beta_3 = 0)$, reporting just the p-value for this test. The gender parameters are jointly significant in Law and Education, and marginally significant in Science and Sport. Ironically, the Faculty with the strongest statistical evidence for gender effects is Law, which is the Faculty with the smallest gender effect in unit SETs.

In our decomposition of the conditional treatment effect in equation (8) the gender of teacher and student affected the expected SET via the formula $\beta_1 - \beta_2 \mathbb{E}[f_s|f_t = 0] + (\beta_2 + \beta_3)\mathbb{E}[f_s|f_t = 1]$. In many of the specifications in Table 3 our estimates of $\beta_2 + \beta_3$ are fairly large (and statistically significant) which is enough to balance or cancel negative estimates of β_1: this is particularly apparent in Education and Science. Conversely the beta coefficients for Business look quite small but that is the Faculty where female teachers get the lowest SETs compared to male teachers. Another important part of the story is that there is a relatively high proportion of SETs completed by female students (i.e. the $\mathbb{E}[f_s|*]$ terms), which may be partly because female students complete SETs more often, as found by Goos and Salomans (2017).

< Table 3 – Gender effects by faculty / subject – about here >

< Table 4 – Fixed effect regressions by faculty / subject – about here >

In Table 4 we report corresponding regressions using fixed effects as described in equations (9) and (10), with the estimates of $\hat{\alpha}_1'$ repeated for ease of comparison. The closest analogue to these results in the published literature is the second specification in Boring (2017, pp.31.ff), although her analysis was based on a multi-section study of seminar teaching rather than lecturing. Her results can be summarized as $\gamma_1 > 0$, $\gamma_2 \leq 0$, for the tutor fixed-effect specification and $\delta_1 > 0$, $\delta_2 \leq 0$, for the student fixed-effect

specification but we find δ_1 negative or close to zero in many sub-samples. As with Table 3, the correspondence between gender effects at student level and gender effects at unit SET level is very weak. For example, within Business male students rate male staff more highly than female staff since γ_1 is positive (and arguably large) but the effect is even larger in Science (where the overall gender effect is smaller) and Education (where the overall gender effect is reversed, since female staff get higher SETs overall).

7. Conclusion and discussion

In this paper we have discussed two important issues about gender effects in Student Evaluations of Teaching.

Our first result was to contrast the unconditional and conditional effects of a teacher being female on their SET score. Unsurprisingly these can be different – phrased as a difference in unconditional and conditional effects the result risks appearing trivial. However, the difference matters because hiring and promotions committees who use SETs will typically only have access to unconditional SETs and published analyses on gender effects of SETs are conditional. In section 3 we provided a framework to discuss the differences between potential discrimination by students in the evaluations and differences in the unit SET scores which might affect hiring or promotion decisions. In section 5 we illustrated that this could be quantitatively important by providing examples from an Italian university.

Our second point was to discuss the usefulness of multi-section studies, i.e. analyses where bias arising from confounding factors can be excluded because the teaching evaluated is of the same material on the same unit. Multi-section studies can usually only be used to evaluate small-group seminar teaching since it is rare for different lecturers to teach the same material to different students on the same unit. It is well known that multi-section studies can potentially identify causes of SETs when students are randomly allocated to teachers, which is plausible. However, a further requirement for multi-section studies is either that teachers never have comparative advantage in teaching different material or that they are randomly allocated to teaching. Neither of these assumptions are appealing.

We conclude with a final comment on the effect of gender biases on hiring and promotion decisions. Gender differences persist, particularly in Economics, despite

discrimination being illegal, but is it plausible that gender effects in SETs play a major part in this under-representation of women? While the answer depends partly upon the magnitude of the gender effect in SETs, the more important issues may be the extent to which women choose to do PhDs in Economics, whether hiring and promotion committees are interested in SETs relative to other factors and the extent to which committee members consciously or unconsciously discriminate against (or for) female applicants. Placing a significant share of the blame on SETs is really an abdication of responsibility by academic staff, conveniently shifting the focus on to students for something over which they have little control.

References

Abrami, Philip C., Sylvia d'Apollonia and Peter A. Cohen (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology,* 82(2), 219–231.

Al-Issa Ahmad and Hana Sulieman (2007). Student evaluations of teaching: perceptions and biasing factors, *Quality Assurance in Education*, 15(3), 302-317.

Babad, Elisha, and Arik Tayeb (2003). Experimental analysis of students' course selection. *British Journal of Educational Psychology*, 73(3), 373-393.

Basow, Susan A., and Suzanne Montgomery (2005). Student Ratings and Professor Self-Ratings of College Teaching: Effects of Gender and Divisional Affiliation. *Journal of Personnel Evaluation in Education*, 18, 91–106.

Bavishi, Anish, Juan M. Madera and Michelle R. Hebl (2010). The Effect of Professor Ethnicity and Gender on Student Evaluations: Judged Before Met. *Journal of Diversity in Higher Education,* 3(4), 245-256.

Bayer, Amanda, and Cecilia Elena Rouse (2016). Diversity in the Economics Profession: A New Attack on an Old Problem. *Journal of Economic Perspectives,* 30(4), 221–242.

Becker, William E., and Michael Watts (1999). How Departments of Economics Evaluate Teaching." *American Economic Review*, 89(2), 344-349.

Becker, William E., William Bosshardt and Michael Watts (2012). How Departments of Economics Evaluate Teaching. *The Journal of Economic Education,* 43(3), 325-333.

Boring, Anne (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics,* 145, 27-41.

Brown, Christopher L., and Stephen M. Kosovich (2015). The impact of professor reputation and section attributes on student course selection. *Research in Higher Education*, 56(5), 496-509.

Bowles, Samuel, and Wendy Carlin (2020). What Students Learn in Economics 101: Time for a Change. *Journal of Economic Literature*, 58(1), 176-214.

Braga, Michela, Marco Paccagnella, and Michele Pellizzari (2014). Evaluating students' evaluations of professors. *Economics of Education Review,* 41, 71-88.

Cannon, Edmund, and Giam Pietro Cipriani (2021). Quantifying halo effects in students' evaluation of teaching. *Assessment & Evaluation in Higher Education*, DOI: 10.1080/02602938.2021.1888868.

Carroll, Scott E., and James E. West (2010). Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors. *Journal of Political Economy*, 118(3), 409-432.

Centra, John A. and Noreen B. Gaubatz (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education*, 70(1), 17-33.

Cipriani, Giam Pietro, and Angelo Zago (2011). Productivity or discrimination? Beauty and the exams. *Oxford Bulletin of Economics and Statistics,* 73(2), 428-447.

Clayson, Dennis E. (2018). Student evaluation of teaching and matters of reliability. *Assessment & Evaluation in Higher Education*, 43(4), 666-681.

Feistauer, Daniela and Tobias Richter (2018). Validity of students' evaluations of teaching: Biasing effects of likability and prior subject interest. *Studies in Educational Evaluation,* 59, 168-178.

Feldman, Kenneth A. (1993). College students' views of male and female college teachers: part II Evidence from students' evaluations of their classroom teachers. *Research in Higher Education*, 34(2), 151-191.

Fenn, Aju Jacob (2015). Student evaluation based indicators of teaching excellence from a highly selective liberal arts college. *International Review of Economics Education*, 18, 11-24.

Goos, Martin and Anna Salomons (2017). Measuring teaching quality in higher education: assessing selection bias in course evaluation. *Research in Higher Education*, 58, 341-364.

He, Jun and Lee A. Freeman (2020). Can we trust teaching evaluations when response rates are not high? Implications from a Monte Carlo simulation, *Studies in Higher Education*, forthcoming.

Hempel, Byron, Kasi Kielbaugh and Paul Blowers (2019). Student evaluation of teaching in an engineering class and comparison of results based on instructor gender. *Engineering Education Research*, 53(2), 91-99.

Kherfi, Samer (2011). Whose Opinion Is It Anyway? Determinants of Participation in Student Evaluation of Teaching. *Journal of Economic Education,* 42(1), 19-30.

Lalla, Michele, and Davide Ferrari (2011). Web-based versus paper-based data collection for the evaluation of teaching activity: empirical evidence from a case study, *Assessment and Evaluation in Higher Education*, 36(3), 347-365.

Linask, Maia, and James Monks (2018). Measuring faculty teaching effectiveness using conditional fixed effects, *Journal of Economic Education,* 49(4), 324-339.

Linse, Angela R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees, *Studies in Educational Evaluation,* 54, 94-106.

Liu, Ou Lydia (2012). Student Evaluation of Instruction: In the New Paradigm of Distance Education. *Research in Higher Education*, 53, 471-486.

May, Ann Mari, Mary G McGarvey and Robert Whaples (2014). Are disagreements among male and female economists marginal at best? A survey of AEA members and their views on economics and economic policy. *Contemporary Economic Policy*, 32(1), 111-132.

McPherson, Michael A., R. Todd Jewell and Myungsup Kim (2009). What Determines Student Evaluation Scores? A Random Effects Analysis of Undergraduate Economics Classes. *Eastern Economic Journal,* 35(1), 37-51.

Mengel, Friederike, Jan Sauerman and Ulf Zölitz (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association,* 17(2), 535-566.

Ragan, James F., and Bhavneet Walia (2010). Differences in Student Evaluations of Principles and Other Economics Courses and the Allocation of Faculty across Courses. *Journal of Economic Education,* 41(4), 335–352.

Rivera, Lauren A., and András Tilcsik (2019). Scaling Down Inequality: Rating Scales, Gender Bias, and the Architecture of Evaluation. *American Sociological Review*, 84(2), 248–274.

Schmidt, Ben (2015). Gendered Language in Teaching Evaluations. Available online at http://benschmidt.org/profGender/#

Spooren, Pieter, Bert Brockx and Dimitri Mortelmans (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, 83(4), 598-642.

Uttl, Bob, Carmela A. White and Daniela Wong-Gonzalez (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, 54, 22-42.

Vanacore, Amalia and Maria Sole Pellegrino (2019). How Reliable are Students' Evaluations of Teaching? A Study to Test Student's Reproducibility and Repeatability. *Social Indicators Research,* 146, 77–89.

Wagner, Natascha, Matthias Rieger and Katherine Voorvelt (2016). Gender, ethnicity and teaching evaluations: evidence from mixed teaching teams. *Economics of Education Review*, 54, 79-94.

Weinberg, Bruce A., Masanori Hashimoto and Belton M. Fleisher (2009). Evaluating teaching in higher education. *Journal of Economic Education*, 40(3), 227-261.

Wolbring, Tobias, and Edgar Treischl (2016). Selection bias in students' evaluation of teaching. *Research in Higher Education*, 57(1), 51-71.

Acknowledgements

We should like to thank university for administrators who provided data to us in an anonymised and useable form and to thank Sarah Smith and Cristina Rossi for comments on a much earlier version of this paper. Any remaining errors are the authors' own.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical considerations

The data used in this study were obtained from an administrative database and anonymised by the university administrators who were not otherwise involved in the data analysis. It was not possible to obtain consent from the respondents and requesting informed consent might have influenced their answers: this issue was considered by the university ethics committee who gave special permission to use the data for this study.

Declaration of interest

None, for either author.

Tables

Table 1 Descriptive statistics

	Number of units	% female teachers	Number of SETs	% female teachers	% female students
All	727	39%	35403	39%	70%
Units with ten or more SETs	557	40%	34527	39%	70%
Attenders	557	40%	27474	39%	70%
Non-attenders	506	41%	7929	39%	71%
L (undergraduate)	331	40%	25866	41%	72%
LM (masters)	191	41%	6168	36%	68%
LM5 (integrated masters)	39	33%	2493	31%	66%
Business	73	29%	4623	34%	52%
Law	60	37%	3208	34%	65%
Languages	90	56%	7383	49%	86%
Humanities	147	37%	6526	33%	68%
Science	72	35%	2300	32%	40%
Education	89	49%	7775	43%	91%
Sport	32	28%	2712	36%	36%

"L" denotes undergraduate students (Laurea Triennale); "LM" denotes masters (Laurea Magistrale); LM5 are five-year degrees. Except for "All", the analysis is confined to units with more than ten SETs. Totals for "L", "LM" and "LM5" are greater than the number of units because some units are shared; similarly, for totals across faculties.

Table 2 Summary statistics of Unit SET scores

	Male teachers			Female teachers			Comparison of means	
	(1) Number of units	(2) Mean	(3) St.Dev.	(4) Number of units	(5) Mean	(6) St.Dev.	(7) Difference	(8) p-value
All	445	3.17	0.43	282	3.18	0.42	0.01	0.72
Units with ten or more SETs	335	3.08	0.39	222	3.12	0.37	0.05	0.16
Attenders	335	3.15	0.41	222	3.19	0.38	0.04	0.20
Non-attenders	301	2.80	0.50	205	2.87	0.48	0.07	0.12
L (undergraduate)	199	3.07	0.36	132	3.07	0.34	0.00	0.99
LM (masters)	112	3.05	0.44	79	3.19	0.41	0.15	0.02*
LM5 (integrated masters)	26	3.27	0.33	13	3.27	0.26	0.00	0.96
Business	52	3.02	0.39	21	2.89	0.35	-0.13	0.17
Law	38	3.20	0.34	22	3.26	0.26	0.05	0.54
Languages	40	2.98	0.46	50	3.13	0.39	0.15	0.10
Humanities	93	3.16	0.34	54	3.24	0.35	0.09	0.14
Science	47	3.03	0.43	25	2.98	0.42	-0.05	0.62
Education	45	2.99	0.40	44	3.11	0.36	0.12	0.14
Sport	23	3.08	0.35	9	2.98	0.21	-0.10	0.43

The unit of observation in the data being analyzed is the mean SET score for each unit. Column (7) reports the difference between columns (5) and (2); column (8) reports the p-value for a two-sided t-test of equality. See note to Table 1 for further explanation of the sub-samples.

Table 3 Gender effects by faculty / subject

	Business	Law	Languages	Humanities	Science	Education	Sport
A: $\bar{y}_t = \alpha'_0 + \alpha'_1 f_t + \eta'_t$							
Female teacher, $\hat{\alpha}'_1$	-0.134	0.052	0.147	0.086	-0.053	0.120	-0.100
	(0.094)	(0.078)	(0.091)	(0.058)	(0.104)	(0.080)	(0.101)
Num. of units	73	60	90	147	72	89	32
B: $y_{t.s} = \alpha_0 + \alpha_1 f_t + \eta_{t.s}$							
Female teacher, $\hat{\alpha}_1$	-0.101	0.058	0.108	0.108+	-0.113	-0.027	-0.047
	(0.080)	(0.105)	(0.113)	(0.058)	(0.115)	(0.090)	(0.106)
Num. of SETs	4623	3208	7383	6526	2300	7775	2712
C: $y_{t.s} = \beta_0 + \beta_1 f_t + \beta_2 f_s + \beta_3 f_t f_s + \boldsymbol{\mu}_1 \boldsymbol{X}_t + \boldsymbol{\mu}_2 \boldsymbol{X}_s + \varepsilon_{t.s}$							
Female teacher, $\hat{\beta}_1$	-0.102	-0.062	0.002	0.037	-0.239+	-0.178*	-0.116
	(0.089)	(0.086)	(0.107)	(0.055)	(0.121)	(0.068)	(0.089)
Female student, $\hat{\beta}_2$	-0.002	-0.089*	-0.065	0.014	-0.155+	-0.068	0.069
	(0.033)	(0.034)	(0.042)	(0.031)	(0.085)	(0.069)	(0.052)
Fem × Fem, $\hat{\beta}_3$	0.073	0.154**	0.059	0.037	0.309*	0.170*	0.061
	(0.049)	(0.049)	(0.052)	(0.051)	(0.128)	(0.083)	(0.070)
Num. of SETs	4535	3191	7217	6437	2254	7589	2701
$\widehat{\beta_1 + \beta_3}$	-0.029	0.092	0.061	0.075	0.07	-0.009	-0.055
	(0.089)	(0.091)	(0.119)	(0.06)	(0.124)	(0.091)	(0.079)
$\widehat{\beta_2 + \beta_3}$	0.070+	0.066+	-0.006	0.052	0.153	0.101*	0.13*
	(0.039)	(0.037)	(0.033)	(0.041)	(0.094)	(0.045)	(0.051)
Joint test (p-value)							
$H_0: \beta_1 = \beta_2 = \beta_3 = 0$	0.284	0.022*	0.480	0.403	0.075+	0.025*	0.070+

Only units with more than ten SETs are analyzed. In panel A an individual observation is a unit, and the dependent variable is the unit SET: $\hat{\alpha}'_1$ corresponds to the differences in Table 2; in panels B and C the individual observations is a SET and the dependent variable is the SET score. All regressions include a constant, not reported to save space; regressions in panel C include control variables: number of students on unit and number of students squared; initial mark of the student on matriculating; type of school that student attended; indicator dummies for attendance and level of degree. The final part of panel C reports the estimates of $\beta_1 + \beta_3$ and $\beta_2 + \beta_3$ derived from the regression in the first part of panel C. $\beta_1 + \beta_3 = 0$ means that female students evaluate male and female teachers the same; $\beta_2 + \beta_3 = 0$ means that female teachers are evaluated the same by male and female students. Heteroskedasticity robust standard errors are reported in parentheses and are clustered at unit level in specifications B and C. The joint test is for all three gender-related parameters being equal to zero and is F(3,N-k): only the p-value is reported. + 0.10 * 0.05 ** 0.01 **

Table 4 Fixed effect regressions by faculty / subject

	Business	Law	Languages	Humanities	Science	Education	Sport
$\bar{y}_t = \alpha'_0 + \alpha'_1 f_t + \eta'_t$							
Female teacher, $\hat{\alpha}'_1$	-0.134	0.052	0.147	0.086	-0.053	0.120	-0.100
	(0.094)	(0.078)	(0.091)	(0.058)	(0.104)	(0.080)	(0.101)
$y_{t,s} = \delta_0 + \delta_1 m_t m_s + \delta_2 f_t f_s + \tau_t + \boldsymbol{\lambda_1 X}_s + \psi_{t,s}$							
Male × Male,	-0.018	0.057+	0.043	-0.031	0.022	0.037	-0.073
$\delta_1 = -\beta_2$	(0.031)	(0.029)	(0.046)	(0.030)	(0.039)	(0.050)	(0.053)
Fem × Fem,	0.086+	0.088**	-0.008	0.078*	0.052	0.083+	0.120*
$\delta_2 = \beta_2 + \beta_3$	(0.047)	(0.030)	(0.041)	(0.038)	(0.103)	(0.044)	(0.050)
Num. of SETs	4535	3191	7217	6437	2254	7589	2701
$y_{t,s} = \gamma_0 + \gamma_1 m_t m_s + \gamma_2 f_t f_s + \sigma_s + \boldsymbol{\lambda_1 X}_t + \phi_{t,s}$							
Male × Male,	0.124**	0.047	0.055	-0.049	0.268***	0.180**	-0.022
$\gamma_1 = -\beta_1$	(0.047)	(0.046)	(0.062)	(0.040)	(0.059)	(0.068)	(0.037)
Fem × Fem,	-0.004	0.165***	-0.018	0.093***	-0.021	0.035	0.050
$\gamma_2 = \beta_1 + \beta_3$	(0.043)	(0.036)	(0.024)	(0.025)	(0.057)	(0.021)	(0.047)
Num. of SETs	3833	2883	6744	6201	2001	7341	2635
Num. of Students	1267	728	1952	1424	624	1697	641

Each column analyzes a separate Faculty, only analyzing units where there were more than ten SETs. The first row of results is reproduced from Table 3 and is reported here to facilitate comparison. Male × Male takes the value one if both teacher and student are male and zero otherwise; Fem × Fem is analogously defined. The first set of regressions contains unit fixed effects and additional controls: initial mark of the student on matriculating; type of school that student attended; where appropriate also an indicator dummy for attendance and indicator dummies for level of degree. The second set of regressions include student fixed effects and additional controls, not reported here to save space: number of students on unit and number of students squared. Notice that in the student-fixed-effect regressions SETs are omitted where a student only submitted one SET. Heteroskedasticity robust standard errors are reported in parentheses. + 0.10 * 0.05 ** 0.01 *** 0.001

Appendix

The appendix contains the following supplementary information.

Tables 3 and 4 in the main paper report regression results using Least Squares (with controls or fixed effects) with the data divided by Faculty. The rational for this is that there seem to be significant inter-Faculty differences. For completeness we report analogous results for the entire University and also consider splitting the sample in other ways (undergraduate versus graduate; attender versus non-attender). These results are reported in Tables A1 and A2 which correspond directly to Tables 3 and 4.

Our analysis has been on the effect of students' responses on unit SET scores and hence we have used Least Squares analysis. Since the individual students' responses are on a four-point Likert scale, a better model might be to use an Ordered Probit analysis. We compare the OLS and Ordered-Probit Analyses in Tables A3 and A4. These results suggest that the estimated gender effects are broadly similar regardless of which method is used.

Sections A5 and A6 provide some additional supporting information about non-attendance and the SET questionnaire.

A1. Gender effects by type of student: University-level analysis

	(1) All	(2) Ten or more	(3) Attenders	(4) Non-attenders	(5) Undergrads	(6) Masters	(7) Int masters
A: $\bar{y}_t = \alpha'_0 + \alpha'_1 f_t + \eta'_t$							
Female teacher, $\hat{\alpha}'_1$	0.012	0.047	0.045	0.070	-0.000	0.146*	0.005
	(0.032)	(0.033)	(0.034)	(0.044)	(0.039)	(0.062)	(0.097)
Num. of units	727	557	557	506	331	191	39
B: $y_{t.s} = \alpha_0 + \alpha_1 f_t + \eta_{t.s}$							
Female teacher, $\hat{\alpha}_1$	0.002	0.006	0.002	0.023	-0.021	0.170*	0.002
	(0.041)	(0.042)	(0.042)	(0.048)	(0.049)	(0.073)	(0.124)
Num. of SETs	35403	34527	26730	7797	25866	6168	2493
C: $y_{t.s} = \beta_0 + \beta_1 f_t + \beta_2 f_s + \beta_3 f_t f_s + \boldsymbol{\mu}_1 \boldsymbol{X}_t + \boldsymbol{\mu}_2 \boldsymbol{X}_s + \varepsilon_{t.s}$							
Female teacher, $\hat{\beta}_1$	-0.059+	-0.055	-0.062	-0.028	-0.078+	0.031	-0.159
	(0.035)	(0.036)	(0.040)	(0.046)	(0.044)	(0.083)	(0.096)
Female student, $\hat{\beta}_2$	-0.033	-0.033	-0.036	-0.022	-0.030	-0.027	-0.081*
	(0.022)	(0.023)	(0.025)	(0.033)	(0.029)	(0.041)	(0.034)
Fem × Fem, $\hat{\beta}_3$	0.105*	0.101*	0.106*	0.079	0.086	0.145+	0.187**
	(0.043)	(0.044)	(0.047)	(0.058)	(0.053)	(0.075)	(0.053)
Num. of SETs	34774	33924	26270	7654	25438	6003	2483
$\widehat{\beta_1 + \beta_3}$	0.045	0.047	0.044	0.052	0.009	0.176*	0.028
	(0.045)	(0.046)	(0.048)	(0.053)	(0.056)	(0.085)	(0.111)
$\widehat{\beta_2 + \beta_3}$	0.072*	0.068*	0.071*	0.057	0.056+	0.117*	0.106*
	(0.028)	(0.029)	(0.031)	(0.043)	(0.033)	(0.056)	(0.04)

Each column analyzes a sub-sample of the data: columns (2)-(7) only analyze units where there were more than ten SETs. In panel A an individual observation is a unit, and the dependent variable is the average SET score; in panels B and C the individual observations is a SET and the dependent variable is the SET score; analogous results for panels B and C using Ordered Probit are reported in the appendix. The regressions in panels A and B include a constant, not reported here to save space; the regression in panel C includes: indicator dummies for faculty; number of students on unit and number of students squared; initial mark of the student on matriculating; type of school that student attended; where appropriate, additional indicator dummies for attendance and level of degree. The final part of panel C reports the estimates of $\beta_1 + \beta_3$ and $\beta_2 + \beta_3$ derived from the regression in the first part of panel C. $\beta_1 + \beta_3 = 0$ means that female students evaluate male and female teachers the same; $\beta_2 + \beta_3 = 0$ means that female teachers are evaluated the same by male and female students. Heteroskedasticity robust standard errors are reported in parentheses and are clustered at unit level in specifications B and C. + 0.10 * 0.05 ** 0.01 *** 0.001

A2. Fixed effect regressions by type of student: University-level analysis

	(1) All	(2) Ten or more	(3) Attenders	(4) Non-attenders	(5) Undergrads	(6) Masters	(7) Int masters
$y_{t,s} = \delta_0 + \delta_1 m_t m_s + \delta_2 f_t f_s + \tau_t + \boldsymbol{\lambda_1 X}_s + \psi_{t,s}$							
Male × Male,	-0.008	-0.007	-0.000	-0.014	-0.006	-0.043	0.051+
$\delta_1 = -\beta_2$	(0.014)	(0.014)	(0.016)	(0.029)	(0.018)	(0.034)	(0.029)
Fem × Fem,	0.070***	0.070***	0.072***	0.070	0.066**	0.068	0.116***
$\delta_2 = \beta_2 + \beta_3$	(0.018)	(0.018)	(0.019)	(0.043)	(0.021)	(0.049)	(0.031)
Num. of SETs	34774	33924	26270	7654	25438	6003	2483
$y_{t,s} = \gamma_0 + \gamma_1 m_t m_s + \gamma_2 f_t f_s + \sigma_s + \boldsymbol{\lambda_1 X}_t + \phi_{t,s}$							
Male × Male,	0.066***	0.059**	0.072***	0.023	0.062**	0.028	0.085
$\gamma_1 = -\beta_1$	(0.018)	(0.019)	(0.022)	(0.046)	(0.020)	(0.061)	(0.053)
Fem × Fem,	0.036**	0.040***	0.037**	0.046	0.006	0.178***	0.127**
$\gamma_2 = \beta_1 + \beta_3$	(0.011)	(0.012)	(0.014)	(0.029)	(0.013)	(0.031)	(0.041)
Num. of SETs	32455	31638	24765	6873	23715	5699	2224
Num. of Students	8358	8333	7639	3682	6155	1619	559

Each column analyzes a sub-sample of the data: columns (2)-(7) only analyze units where there were more than ten SETs. Male × Male takes the value one if both teacher and student are male and zero otherwise; Fem × Fem is analogously defined. The first set of regressions contains unit fixed effects and additional controls: initial mark of the student on matriculating; type of school that student attended; where appropriate also an indicator dummy for attendance and indicator dummies for level of degree. The second set of regressions include student fixed effects and additional controls, not reported here to save space: number of students on unit and number of students squared. Notice that in the student-fixed-effect regressions SETs are omitted where a student only submitted one SET. Heteroskedasticity robust standard errors are reported in parentheses. + 0.10 * 0.05 ** 0.01 *** 0.001

A3. Comparison of Least Squares and Ordered Probit estimation: regressions by type of student

	(1) All	(2) Ten or more	(3) Attenders	(4) Non-attenders	(5) Undergrads	(6) Masters	(7) Int masters
Least Squares							
Female teacher, $\hat{\beta}_1$	-0.059+	-0.055	-0.062	-0.028	-0.078+	0.031	-0.159
	(0.035)	(0.036)	(0.040)	(0.046)	(0.044)	(0.083)	(0.096)
Female student, $\hat{\beta}_2$	-0.033	-0.033	-0.036	-0.022	-0.030	-0.027	-0.081*
	(0.022)	(0.023)	(0.025)	(0.033)	(0.029)	(0.041)	(0.034)
Fem × Fem, $\hat{\beta}_3$	0.105*	0.101*	0.106*	0.079	0.086	0.145+	0.187**
	(0.043)	(0.044)	(0.047)	(0.058)	(0.053)	(0.075)	(0.053)
Num. of SETs	34774	33924	26270	7654	25438	6003	2483
$\widehat{\beta_1 + \beta_3}$	0.045	0.047	0.044	0.052	0.009	0.176*	0.028
	(0.045)	(0.046)	(0.048)	(0.053)	(0.056)	(0.085)	(0.111)
$\widehat{\beta_2 + \beta_3}$	0.072*	0.068*	0.071*	0.057	0.056+	0.117*	0.106*
	(0.028)	(0.029)	(0.031)	(0.043)	(0.033)	(0.056)	(0.04)
Ordered Probit							
Female teacher, $\hat{\beta}_1$	-0.085+	-0.078	-0.092+	-0.032	-0.110+	0.047	-0.266+
	(0.049)	(0.050)	(0.056)	(0.059)	(0.061)	(0.109)	(0.142)
Female student, $\hat{\beta}_2$	-0.054+	-0.053+	-0.060+	-0.031	-0.048	-0.043	-0.133**
	(0.030)	(0.031)	(0.034)	(0.042)	(0.040)	(0.053)	(0.051)
Fem × Fem, $\hat{\beta}_3$	0.147*	0.142*	0.154*	0.096	0.121+	0.184+	0.297***
	(0.057)	(0.059)	(0.064)	(0.073)	(0.072)	(0.097)	(0.081)
Num. of SETs	34774	33924	26270	7654	25438	6003	2483
$\widehat{\beta_1 + \beta_3}$	0.062	0.064	0.062	0.063	0.011	0.231*	0.031
	(0.06)	(0.061)	(0.065)	(0.067)	(0.075)	(0.108)	(0.16)
$\widehat{\beta_2 + \beta_3}$	0.093*	0.089*	0.094*	0.064	0.072	0.141+	0.164**
	(0.038)	(0.038)	(0.042)	(0.055)	(0.045)	(0.074)	(0.061)

This table corresponds to Table 3 in the main text and the Least Squares results are repeated here to facilitate comparison with the Ordered Probit results. Both specifications have additional controls as described in Table 3. The Ordered Probit model specifies that the explanatory variables determine a latent variable which then determines the choice of 1,2,3 or 4 in the evaluation: the parameter estimates here show the relationship between the relevant three explanatory variables and the latent variable. Standard errors are reported in parentheses and are clustered at unit level. + 0.10 * 0.05 ** 0.01 *** 0.001

A4. Comparison of Least Squares and Ordered Probit estimation: regressions by faculty / subject

	(1) Business	(2) Law	(3) Languages	(4) Humanities	(5) Science	(6) Education	(7) Sport
Least Squares							
Female teacher, $\hat{\beta}_1$	-0.102	-0.062	0.002	0.037	-0.239+	-0.178*	-0.116
	(0.089)	(0.086)	(0.107)	(0.055)	(0.121)	(0.068)	(0.089)
Female student, $\hat{\beta}_2$	-0.002	-0.089*	-0.065	0.014	-0.155+	-0.068	0.069
	(0.033)	(0.034)	(0.042)	(0.031)	(0.085)	(0.069)	(0.052)
Fem × Fem, $\hat{\beta}_3$	0.073	0.154**	0.059	0.037	0.309*	0.170*	0.061
	(0.049)	(0.049)	(0.052)	(0.051)	(0.128)	(0.083)	(0.070)
Num. of SETs	4535	3191	7217	6437	2254	7589	2701
$\widehat{\beta_1 + \beta_3}$	-0.029	0.092	0.061	0.075	0.07	-0.009	-0.055
	(0.089)	(0.091)	(0.119)	(0.06)	(0.124)	(0.091)	(0.079)
$\widehat{\beta_2 + \beta_3}$	0.07	0.066	-0.006	0.052	0.153	0.101*	0.13*
	(0.039)	(0.037)	(0.033)	(0.041)	(0.094)	(0.045)	(0.051)
Ordered Probit							
Female teacher, $\hat{\beta}_1$	-0.138	-0.113	0.010	0.053	-0.323*	-0.252*	-0.174
	(0.123)	(0.132)	(0.145)	(0.078)	(0.159)	(0.100)	(0.123)
Female student, $\hat{\beta}_2$	-0.015	-0.142**	-0.095+	0.017	-0.220*	-0.096	0.085
	(0.046)	(0.048)	(0.056)	(0.044)	(0.107)	(0.094)	(0.075)
Fem × Fem, $\hat{\beta}_3$	0.101	0.243***	0.078	0.049	0.405*	0.233*	0.098
	(0.069)	(0.072)	(0.070)	(0.073)	(0.161)	(0.112)	(0.100)
Num. of SETs	4535	3191	7217	6437	2254	7589	2701
$\widehat{\beta_1 + \beta_3}$	-0.037	0.13	0.089	0.102	0.081	-0.019	-0.076
	(0.126)	(0.136)	(0.155)	(0.085)	(0.155)	(0.123)	(0.113)
$\widehat{\beta_2 + \beta_3}$	0.086	0.102+	-0.016	0.067	0.185	0.137*	0.183*
	(0.054)	(0.056)	(0.045)	(0.06)	(0.121)	(0.064)	(0.073)

This table corresponds to Table 4 in the main text and the Least Squares results are repeated here to facilitate comparison with the Ordered Probit results. Both specifications have additional controls as described in Table 4. The Ordered Probit model specifies that the explanatory variables determine a latent variable which then determines the choice of 1,2,3 or 4 in the evaluation: the parameter estimates here show the relationship between the relevant three explanatory variables and the latent variable. Standard errors are reported in parentheses and are clustered at unit level. + 0.10 * 0.05 ** 0.01 *** 0.001

A5. Self-reported reasons for non-attendance at lectures

In the SET students initially had to declare themselves to be attenders or non-attenders (*frequentanti* or *non-frequentanti*) where non-attendance included both complete non-attendance and rare attendance. Non-attenders were then asked: Specify the main reason for your non- or limited attendance (*Indicare il motivo principale della non frequenza o della frequenza ridotta alle lezioni*) and had to choose from the following possible responses:

Response	**Proportion of responses**
Attendance at other courses *Frequenza lezioni di altri insegnamenti*	22%
Attendance not particularly useful to prepare for the exam *Frequenza poco utile ai fini della preparazione dell'esame*	7%
The facilities dedicated to the teaching activities of this course do not allow all interested students to participate *Le strutture dedicate all'attività didattica non consentono la frequenza agli studenti interessati*	1%
Work *Lavoro*	41%
Other reasons *Altro*	29%

A6. Additional information about the evaluation questionnaire

The SET data are administrative data that were collected by the university registry and made available to the researchers in anonymized form after matching to other variables.

Each student is required to take a separate online Student Evaluation of Teaching for each unit (or module). Until the student has completed the SET, they are not allowed to take the examination. As with many Italian universities, some students are enrolled who neither attend lectures nor take examinations: similarly, some students who are enrolled do not take the SET. The issue of enrolled students not engaging with their courses in the Italian university system is discussed by Cipriani and Zago (2011).

Of the students that do take the SET the first question required students to specify whether they were attenders or non-attenders: this then determined the precise list of questions that they received subsequently, although many questions were the same. Attenders were asked why they attended the course and non-attenders were asked why they did not attend.

In a parallel piece of analysis, we are analysing the relationship between answers to different questions. Consistent with findings of "halo effects" the correlations between responses to different questions are very high. In this paper we only analyze the final question for the overall evaluation of the unit: for both attenders and non-attenders this was:

- On the whole, are you satisfied with the organisation and teaching of this course?
- *È complessivamente soddisfatto/a di come è stato svolto questo insegnamento?*