Working Paper
No. 617

Gender, ethnicity and teaching evaluations
Evidence from mixed teaching teams

Natascha Wagner, Matthias Rieger, Katherine Voorvelt

March 2016

ISSN 0921-0210

The International Institute of Social Studies (ISS) is Europe's longest-established centre of higher education and research in development studies. On 1 July 2009, it became a University Institute of the Erasmus University Rotterdam (EUR). Post-graduate teaching programmes range from six-week diploma courses to the PhD programme. Research at ISS is fundamental in the sense of laying a scientific basis for the formulation of appropriate development policies. The academic work of ISS is disseminated in the form of books, journal articles, teaching texts, monographs and working papers. The Working Paper series provides a forum for work in progress which seeks to elicit comments and generate discussion. The series includes academic research by staff, PhD participants and visiting fellows, and award-winning research papers by graduate students.

Working Papers are available in electronic format at www.iss.nl

Please address comments and/or queries for information to:

International Institute of Social Studies
P.O. Box 29776
2502 LT The Hague
The Netherlands

Or

E-mail: wpapers@iss.nl

Table of Contents

Abstract

This paper studies the effect of teacher gender and ethnicity on student evaluations of teaching quality at university. We* analyze a unique data-set featuring mixed teaching teams and a diverse, multicultural, multi-ethnic group of students and teachers. Co-teaching allows us to study the impact of teacher gender and ethnicity on students' evaluations of teaching exploiting within course variation in an empirical model with course-year fixed effects. We document a negative effect of being a female teacher on student evaluations of teaching, which amounts to roughly one fourth of the sample standard deviation of teaching scores. Overall women are 11 percentage points less likely to attain the teaching evaluation cut-off for promotion to associate professor. The effect is robust to a host of co-variates such as course leadership, teacher experience and research quality. There is no evidence of a corresponding ethnicity effect. Our results point to an important gender bias and indicate that the use of teaching evaluations in hiring and promotion decisions may put female lectures at a disadvantage.

Keywords

Student evaluations of teaching, gender, ethnicity, bias, course fixed effects.

JEL classification

I21, J71.

* We wish to thank the seminar participants of the 37th EAIR forum in Krems, Austria and our colleagues at the International Institute of Social Studies for their constructive feedback. In particular Arjun Bedi, Peter van Bergeijk, Jeff Handmaker, Helen Hintjens, Freek Schiphorst, Mindi Schneider, Karin Astrid Siegmann, Nynke-Jo Smit and Irene van Staveren. All remaining errors are our own.

The authors: Natascha Wagner, Matthias Rieger, Katherine Voorvelt, International Institute of Social Studies of Erasmus University Rotterdam, The Hague. Corresponding author: wagner@iss.nl

1 INTRODUCTION

We study the impact of teacher gender and ethnicity on student evaluations of teaching (SET) at a Dutch university using a novel identification strategy exploiting within course variation. SETs are meant to reflect the effectiveness of the teacher in delivering course material in higher education institutions. SETs are used to measure course quality as perceived by students and have been widely implemented for almost hundred years now (Carrell and West, 2010; Marsh, 1984; Guthrie, 1954). Yet, controversies about the content and quality of student evaluations of teaching are almost as old as the teaching evaluations themselves (Marsh, 1984; Cadwel and Jenkins, 1985; Marsh and Groves, 1987; Abrami and d'Apollonia, 1991; Marsh, 1991).

Existing research does suggest that the resulting average evaluations are reliable and stable, but to a large extent a function of teacher characteristics and behavior rather than course content and quality *per se* (Pounder, 2007; Marsh, 1987). Whether student evaluations of teaching are related to course grades and workload is contested.[1] At the same time, only a very small positive association between teaching evaluations and student learning is found (see Beleche et al., 2012).[2] Apart from these caveats, the use of average SET scores ignores issues related to response rates and response variability (Stark and Freishtat, 2014). The notion that assessments in general tend to reflect on contextual factors and often on gender rather than exclusively dealing with the subject matter is further reinforced by the Harvard Implicit Association Test (Greenwald et al., 1998), which revealed implicit bias against women in positions of power (Crockett, 2015; Mo, 2014).

[1] While Marsh and Roche (2000) argue it is not, there is increasing evidence that teachers who give higher grades also receive better evaluations (Ewing, 2012; Carrell and West, 2010; Weinberg et al., 2009; Langbein, 2008; Isely and Singh, 2005; Johnson, 2003; Krautmann and Sander, 1999).

[2] Braga et al. (2014) show that students' evaluations of teachers are negatively correlated with a more objective measure of teaching effectiveness and quality, which is student performance in subsequent coursework. Becker and Watts (1999) assess data from a survey among economics departments in the US and show that student evaluations of teaching explain less than 50 percent of the variation in student learning outcomes. SET scores are not very highly correlated with other measures of good teaching such as peer review. Moreover, students seem to give a beauty premium to their professors (Hamermesh and Parker, 2005). Other than teacher characteristics, situational factors, such as whether the faculty association or the student association are in charge of organizing the evaluation also influence the outcome of the evaluations (Abrami et al., 1976).

At the same time, the causal determinants of SETs are still poorly understood. Such an understanding is important since SETs are increasingly used to inform promotions and may impact negatively on the academic careers of young faculty members in higher education (Boring et al., 2016; Walstad and Saunders, 1998; Seldin, 1993).

In this paper we focus on two important teacher traits and their potentially discriminatory impact on SETs: gender and ethnicity. In particular, the issue of gender perceptions and bias in academia has received considerable public attention (Hay, 2016; Kamenetz, 2016; Poropat, 2014). Hay (2016) underlines that women in academia are expected to be nice, caring and good-looking. Depending on their age female professors are seen as "girlfriend" or "mother" and not necessarily as professionals. Boring et al. (2016) show that gender bias is even found in objective aspects of teaching and varies by discipline and student gender. The study documents double standards applied to male and female teachers both in the United States and in France. In addition to gender we assess the performance of teachers from ethnic minorities in student evaluations of teaching since these two traits tend to coincide. While more and more women and teachers of different ethnic backgrounds enter academia, white male professors are still the norm and tend to achieve better teaching evaluations (Boring et al., 2016; Basow and Silberg, 1987).

We contribute a new causal identification strategy to assess the impact of teacher traits on student evaluations of teaching. Average scores differ by subject and a naïve analysis where one combines all courses and therefore cannot reveal gender-differences as suggested by Schmidt (2015).[3] We make use of a study setting where most lectures teach more than one course and where many courses are co-taught by mixed gender and ethnicity teams. This allows us to study the impact of gender and ethnicity on student evaluations *within* the same course. This strategy controls for course heterogeneity and for self-selection of teachers and students into courses, all of which are determinants of evaluations (Schmidt, 2015; Ongeri, 2009; Cashin, 1990). We document significantly lower grades in teaching evaluations for women, but

[3] Schmidt (2015) makes use of an online search tool for words and phrases and applied this to web ratings of professors in about 14 million reviews from RateMyProfessor.com. He considers widely used terms to describe male and female teachers. Across academic disciplines, men are far more likely to be considered funny. And not only that, they are more likely to be considered brilliant and a genius, whereas women are more likely to be rated annoying, strict and harsh. In line with gendered stereotyping women are more likely to be judged nice, helpful and friendly. Women are also more likely to be rated incompetent.

only once we control for such course unobservables. In other words teacher evaluations are not gender blind, and gender effects explain roughly one fourth of their sample standard deviation. Our findings suggest that women are 11 percentage points less likely to attain the teaching evaluation cut-off for promotion to associate professor. The gender effect is also important in magnitude compared to other significant determinants of SETs such as teacher quality as measured by the number of top publications per year. Female teachers would need a sizeable 4.79 A publications (the sample average of A publications per year is 0.86) to offset the negative direct gender impact on the student evaluations of teaching. In contrast, we do not find evidence of an ethnicity effect in the evaluations and attribute this finding to the multi-ethnic student pool. Our main result and its magnitude are in line with an online experiment with 43 students by MacNell et al. (2014). The crux of the experiment is that the students never saw or heard their teacher because of the online format of the course. The supposedly "male" teacher received higher grades, regardless of the actual gender.[4]

Interestingly, we find that the negative gender effect is reversed in the major for gender studies and social justice suggesting that students can be sensitized to gender issues. Finally, we cross-validate our main findings by looking at the effect of teaching team composition on overall evaluations for the course. While gender matters for individual evaluation of teachers, the share of female teachers has no effect on how students perceive the course in general. We take this as additional evidence of personal discrimination.

The remainder of the paper is organized as follows. In Section 2 we describe the setting of our study. The empirical approach is introduced in Section 3 and our dataset in Section 4. Section 5 presents the results and assessment of their robustness. Section 6 concludes with a brief discussion.

2 STUDY SETTING

Our study is set in a unique, multi-ethnic institute where awareness for social and gender-justice is one of the teaching missions. This makes it a particularly interesting case to ascertain whether student evaluations of teaching are sensitive to gender and ethnicity. The International Institute

[4] Related, Bachen et al. (1999) argues that the gender of both the teacher and the student, as well as their interaction, are associated with the resulting scores in the student evaluation of teaching.

of Social Studies (ISS) of Erasmus University is a graduate school in The Netherlands that brings together students and teachers from the Global South and North.[5]

ISS is one of the oldest and largest centers in Europe for the comparative study and research of social, political and economic development. Until today, most of the students come from developing countries and countries in transition and the teaching staff is comprised of a diverse group. In the academic year 2014/15, the MA program was comprised of more than 150 MA students from 46 different countries and 55 teaching staff members from more than 15 different nations. The institute currently offers five majors, namely (i) Agrarian, Food and Environmental Studies, (ii) Economics of Development, (iii) Governance, Policy, and Political Economy, (iv) Human Rights, Gender and Conflict Studies: Social Justice Perspectives and (v) Social Policy for Development.[6]

3 EMPIRICAL STRATEGY

The empirical challenge for any research on the impact of gender and ethnicity on SETs is that it is hard to account for self-selection into certain types of courses. The type of course is potentially correlated with the gender and ethnic background of the teacher. If, for instance, female teachers select courses where they expect to get higher teaching grades, then the simple difference in means between male and female teaching scores will be biased towards zero.

Ideally one should thus study teaching evaluations by the same students in the same course for female and male teacher in order to address such concerns. In other words, we would like to keep the type of course and content fixed when we look at gender and ethnicity effects. With

[5] ISS was founded as an independent institute in 1952 when decolonization had been set in motion in India, Pakistan, Ceylon and then Indonesia. Initially, the Dutch government considered that ISS could influence the former colonies by training bureaucrats, policy and decision makers. The Institute has trained and influenced the thinking of future policy-makers across the globe. Since 2009, it functions as a University Institute *sui generis* within the Erasmus University of Rotterdam (EUR). The overall mission of ISS is to be a research-led graduate school in social sciences that is teaching-based, contributes to public debates and influences public opinion and policy-making on issues of development, equity and social justice worldwide.

[6] ISS offers a 15.5 months MA program and a four-year PhD program in Development Studies and several postgraduate diploma programs and tailor-made short courses. The ISS master's program is policy-oriented and combines different strands of modern social sciences stretching from political sciences to youth and gender studies, as well as human rights, environment, and economics.

our data we can do exactly this: Most courses at the Institute of Social Studies are co-taught, often by female and male teaching staff. We have a sample of 688 course-teacher observations for five consecutive years of teaching starting in the academic year 2010-11. Almost 45 percent of the observed teacher evaluations in our sample pertain to women, 75 percent of the courses are co-taught and more than 65 percent are co-taught by mixed gender or female teaching teams.

Suppose we estimate the following linear regression model:

$$Teaching\_score_{ict} = \beta_0 + \beta_f\, female_{ict} + \beta_{N\text{-}C}\, Non\_Caucasion_{ict} + \varepsilon_{ict} \qquad (1)$$

where the outcome variable $Teaching\_score_{ict}$ corresponds to the teaching evaluation teacher *i* has gotten for course *c* in year *t*. The dummy variable $female_{ict}$ takes on a value of one for female teachers and zero otherwise. The dummy variable $Non\_Caucasion_{ict}$ measures ethnic origin coding one for Non-White teachers and zero otherwise.

We are interested in the coefficients β_f and $\beta_{N\text{-}C}$ associated with the teacher's gender and ethnicity in equation (1). However note that the error term ε_{ict} in the above specification is likely correlated with our variables of interest due to sorting into certain types of courses.

We address the endogeneity issues related to equation (1) in four ways. First, we add course-specific fixed effects λ_c. Second, to account for overall changes in student evaluations of teaching we further include time specific effects t_t. Third, we also combine these two fixed effects in the most demanding econometric specification and incorporate course-year-fixed effects allowing us to capture course specific heterogeneity in a certain year. Finally, we control for other observable course and teacher characteristics to gauge the sensitivity to omitted variables specific to the teacher. These characteristics include the number of student participants, the proportion of students evaluating the course and whether the observed teacher is the course leader, as well as the teacher's publications. For a subsample of teachers, for whom we have detailed age information, we can also control for experience as captured by age.

One remaining worry is that certain male teachers might prefer to teach with other male teachers or alone, and that this preference is related to overall course characteristics. In other words, the sub-sample of courses featuring mixed teaching might be "special" and our results would have reduced external validity. To investigate this concern we perform "placebo" tests by regressing overall course evaluations and characteristics (such as student workload) on the gender

composition of the teaching team. As we report below, the gender composition of the teaching team is unrelated to these overall course outcomes. In other words, gender only matters for individual assessments of teaching, which is suggestive of a gender bias.

After netting out course and teacher characteristics, we argue that remaining effects associated with gender and ethnicity are suggestive of discrimination. Standard errors are clustered at the level of fixed effects.

4 DATA

Student evaluations of teaching at ISS have been based on the same questionnaire across courses over a five year period. This results in a dataset with more than 650 comparable teaching evaluations for a total of 272 courses. Table 1 presents the number of courses offered per academic year and the number of teachers giving them. Over time we observe a steady decline in the number of courses due to a reorganization of courses and programs. We also observe a decrease in the number of teachers from 62 in 2010-11 to 55 in 2014-15 due to human restructuring. We capture these time trends using time dummies.

Next we turn to the descriptive statistics presented in Table 2. The average teaching score is 4.27 on a 1 to 5 Likert scale indicating that the courses are well perceived by the students. Note that each student anonymously fills in one evaluation form featuring questions about the course in general and one question about each specific teacher. The average score reflects the perceived teaching quality of each teacher. Out of the 688 course evaluations there are 11 that obtained the maximum of 5 points but there is also 1 course with a score as low as 1.82 points. There are 8 courses that received an average score below 3. The average and the median course grade are identical indicating that the descriptive statistics are not driven by extreme values. The one-standard deviation window around the mean ranges from 3.83 to 4.71 suggesting that already small differences in the student evaluations of teaching are decisive for passing or failing the evaluation cut-off point of 4, which needs to be reached by a teacher in order to obtain tenure at ISS.

Year	Number of courses	Number of teachers
2010-11	73	62
2011-12	60	52
2012-13	53	58
2013-14	37	58
2014-15	49	55
Average	54.40	57.00

Table 1: Number of courses and teachers across the years.

We further collected information on observable course characteristics: Almost half the evaluations, namely 44.19 percent, are for female teachers. Moreover, more than one third of the observations are for teachers of Non-Caucasian background, underlining the ethnic diversity of the teaching staff. The student response rate to the evaluations of teaching is as high as 87.0 percent. This high response rate is driven by an incentive scheme that releases exam grades late for students who did not complete the course evaluation. The average course consists of roughly 32 students. Slightly less than one third of the observed teacher evaluations are for course leaders, reflecting again that the majority of the courses are co-taught. Course leaders are in charge of organizing the overall course and the main contact person for students on administrative matters. We only have information about the age of the teaching staff for 599 of the 688 observations, i.e. 63 of the 93 teachers in the sample. In that sub-sample the average age of the teacher is about 48 years.

In our empirical specification we will also include interaction terms between gender and other covariates to identify through which channels teaching scores can be affected. For instance, we interact the ethnicity and the gender dummies since 16.57 percent of the total observations are for Non-White, female teachers. This highlights that roughly half of the course evaluations for Non-White teachers are for female teachers and that the pool of teachers in our study is truly heterogeneous along the ethnic and gender dimension. We further interact course leadership with the gender of the teacher. Compared to men, women are less likely to be course leaders. Women account for only a third of the course leaderships.

We also control for teacher quality in our empirical specification. Since ISS is a research led higher education institute we use the number of research points a teacher collects in any given year. These points are calculated based on a journal classification by the European Association of Development Research and Training Institutes (EADI). We further explain the content and set-up of the EADI ranking below when we discuss the related findings. The average teacher has 10.72 research points in any given year. The top 33% publications within a research area based on the Thomson Reuter Journal Citation Reports (e.g. economics, political science etc.) are rated as A publications. On average 0.86 A publications are produced per teacher and year.

We further compare the average characteristics of the two sub-samples of female and male teachers (Table 2). When considering a simple average we do not find any statistically significant differences in course evaluations for women and men. The share of Non-Caucasian male and female teachers is also statistically equal, as is the number of course participants across gender and the response rate to the evaluations. However, women are significantly less likely to be course leaders with the difference being statistically significant at the 1 percent level. Women also tend to be younger. While this difference is statistically significant, it amounts to only 1.52 years. Moreover, women tend to have 3.77 EADI research points less. However, when comparing the sum of A publications, women and men fare equally well. A difference emerges in B publications; men tend to have 0.72 B publications more per year. There is no gender difference in C publications.

These simple differences underline the need to account for observable teacher und course characteristics in our empirical analysis. The picture looks similar when splitting the sample along the ethnic dimension. Here we do not observe statistically significant differences in any of the raw means except for the number of A publications. Non-Caucasian teachers tend to have 0.26 A publications more per year with the difference being statistically significant at the 5 percent level. The results are not presented for the sake of brevity but of course available on request from the authors.

Variable	Total				Female teachers		Male teachers		Diff. in means
	Mean	Std.Dev.	Min	Max	Mean	Std.Dev.	Mean	Std.Dev.	*p*-value
Teaching grade	4.271	0.443	1.82	5	4.255	0.471	4.284	0.419	0.389
Observable teacher and course-related characteristics									
Female teacher	0.442		0	1					
Non-Caucasian teacher	0.359		0	1	0.375		0.346		0.437
Number of participants	32.041	34.538	3	187	31.250	32.529	32.667	36.079	0.594
Response rate to teaching evaluation	0.870	0.137	0.417	1.385	0.871	0.144	0.870	0.132	0.984
Course leader	0.327		0	1	0.260		0.380		0.001***
Age	48.170	9.253	29	65	47.386	9.083	48.906	9.364	0.044**
Teacher quality as captured by research output									
EADI research points	10.716	9.657	0	64	8.661	7.700	12.432	10.744	0.000***
Number of A publications	0.862	1.245	0	7	0.811	1.037	0.904	1.395	0.402
Number of B publications	1.531	1.921	0	15	1.137	1.288	1.860	2.272	0.000***
Number of C publications	1.110	1.377	0	6	0.912	1.408	1.276	1.331	0.003
Interaction terms									
Non-Caucasian teacher × female	0.166		0	1	0.375				
Course leader × female	0.115		0	1	0.260				

Table 2: Descriptive statistics. The unit of observation is student evaluations of teaching. The total number of observations is 688 of which 304 correspond to evaluations of female teaching. For the variable age we have 599 observations of which 290 correspond to evaluations of female teaching. For the variables EADI points and number of A, B and C publications we have 499 observations, of which 227 correspond to evaluations of female teaching.

5 RESULTS

5.1 MAIN RESULTS

Is there a gender difference in students' rating of teachers? Based on the simple comparison of means, the answer is no. However a simple difference in means is biased towards zero as we show now.

In Table 3, we present findings from an OLS and various fixed effect models. Column 1 reveals that jointly controlling for gender and ethnicity in teaching grades suggests that female and

Non-White teachers get lower grades as both coefficient estimates are negative. However the estimates are not statistically significant and economically small.

Dependent variable: Teaching score	(1) OLS	(2) FE	(3) FE	(4) FE
Female teacher	-0.029	-0.122*	-0.121*	-0.125***
	(0.054)	(0.064)	(0.062)	(0.048)
Non-Caucasian teacher	-0.021	0.067	0.062	0.047
	(0.049)	(0.048)	(0.046)	(0.038)
Observations	688	688	688	688
Course fixed effects [Number]	no	yes [95]	yes [95]	no
Year fixed effects [Number]	no	no	yes [4]	no
Course-year fixed effects [Number]	no	no	no	yes [272]

Table 3: Main results. Standard errors are clustered at the course level in specifications 1 to 3 and at the course-year level in specification 4.
*/ **/ *** p < 0.10/0.05/0.01, respectively.

In this simple OLS model we are ignoring that teachers and students self-select into courses. Failing to control for course characteristics leads to attenuation bias. In other words, course characteristics are correlated with the gender of teachers and bias the correlation to zero. Needless to say that teaching grades from a statistics and a history course are incommensurable since course contents and audience are very different.

In Column 2 of Table 3, we account for course-specific effects and find that female teachers receive considerably lower teaching grades. The coefficient estimate of -0.12 seems to suggest a relatively small effect at first sight; it corresponds to only 2.86 percent of the average grade. However taking into account that the average teaching score of 4.27 is rather high and the distribution is very tight around that mean as indicated by the standard deviation of 0.44, the estimated effect is substantial. It explains 27.55 percent of the sample standard deviation. The magnitude of our result is in line with an online experiment by MacNell et al. (2014). The study reports an effect of 0.15 on a similar five-point Likert scale.

More importantly, the gender effect could have sizable repercussions on promotion decisions. The cut-off for promotion to associate professor at ISS is a teaching score of at least 4. We re-ran specification 2 of Table 3 with a binary dependent variable that takes on the value of 1 if

the cut-off is reached and zero otherwise. Women are 11 percentage points less likely to attain this promotion cut-off compared to men teaching within the same course (*p*-value= 0.040).

The coefficient associated with being Non-Caucasian has changed sign but remains statistically insignificant suggesting that along the ethnic dimension there is no bias. To assess whether these findings are robust, we further control for changes over time (Table 3, Column 3). When including time fixed effects together with the course fixed effects the coefficient estimates remain virtually unchanged. In other words, gender effects are stable over the years.

Finally, our most demanding specification exploits within course-year variation (Table 3, Column 4). Comparing grades between female and male teacher within the same course environment and the same group of students allows us to account for self-selection into the type of courses and overall course characteristics. Again, we find gender but no ethnicity effects. The coefficient associated with teacher gender is marginally bigger in absolute terms, and more precisely estimated than in the two preceding fixed effect models. Across fixed effect models we coherently document gender effects with the coefficient not being sensitive to the type of fixed effects we employ.

5.2 GENDER EFFECTS AND THE INCLUSION OF CO-VARIATES

Next we examine the sensitivity of the gender effect to controlling for observable characteristics of courses and teachers. We argue that these conditional fixed effect estimates are suggestive of gender bias.

Do student attendance and response rates matter? In Column 1 of Table 4 we include the number of participants and the response rate to the teaching evaluation. The remaining effect after netting out observable course characteristics can be more plausibly attributed to gender discrimination. Since these two variables are identical within a course for a given year, we employ course and year-fixed effects independently. As can be seen, only the response rate, but not the number of participants is correlated with the teaching grade. It suggest that a higher response rate is better for the teachers as it averages out extreme responses. Most importantly, despite controlling for the number of course participants and the response rate to the teaching evaluations, we still find a robust and negative gender effect.

Does course leadership matter? In Column 2 of Table 4 we account for course leadership in addition to the class characteristics. Course leaders tend to be significantly higher evaluated as compared to non-leaders. The coefficient of 0.150 associated with course leadership is slightly bigger in absolute terms as compared to the negative gender coefficient of 0.107. In other words, women could potentially "offset" the negative gender effect by being course leaders. We run a robustness check on the course leader effect with course-year fixed effects in Column 3 of Table 4. The coefficient associate with course leadership drops slightly in size but is still statistically significant at the 1 percent level. Also, the negative gender effect persists. We directly test the offsetting effect for female course leaders in Column 4 of Table 4 by including an interaction term between gender and course leadership. In terms of size, the coefficient estimate associated with the course leader-gender interaction offsets the negative impact of gender: statistical equality of the absolute values of these two coefficient estimates cannot be rejected (*p*-value= 0.733). Once we include the interaction term associated with course leadership the direct effect of course leadership disappears with all the positive effect of course leadership applying to female course leaders. Therefore, we proceed by further studying the channels for gender bias and interact both gender and course leadership, as well as gender and ethnicity (Table 4, Column 5). The coefficients associated with ethnicity and the interaction of ethnicity and gender remain insignificant. Again, the direct negative gender effect and the positive effect for female course leaders persist.

These findings can be interpreted in the sense that in the context of co-teaching women can improve their course evaluations by taking on course leadership. One has to keep in mind, however, that this is only a partial remedy to the gender bias in student evaluations of teaching as course leadership comes with an additional workload.

Age is another possible source for discrimination and also a proxy of experience and teaching quality. Marsh (2007) makes use of 13 years of evaluations for 195 courses and finds little evidence that teacher effectiveness changes with age. We only have age information for a subsample of 63 of the 93 observed teachers, resulting in 599 observations of SETs. When estimating the model for the sub-sample including age we find a significant, inverse U-related relationship between age and teaching scores. The turning point is age 45. This roughly corresponds to the time when academics have obtained full professorship suggesting some small decline in teaching quality once individuals get tenured. Including age in the specification does not alter the negative and statistically significant gender impact. To assure that the smaller

sample for the age estimations is not biased we re-estimate the basic model with course-year fixed effects and only the gender and ethnicity dummy. We obtain the negative impact on gender and again no impact on ethnicity. The negative coefficient associated with gender is in line with our main estimates.

Dependent variable: Teaching score	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female teacher	-0.123*	-0.107*	-0.111**	-0.160***	-0.176***	-0.091*	-0.097*
	(0.062)	(0.059)	(0.046)	(0.054)	(0.061)	(0.048)	(0.051)
Non-Caucasian teacher	0.062	0.054	0.041	0.053	0.031	0.055	0.050
	(0.046)	(0.046)	(0.039)	(0.040)	(0.055)	(0.041)	(0.042)
Non-Caucasian teacher × female					0.043		
					(0.076)		
Number of participants (log)	-0.008	-0.009					
	(0.093)	(0.093)					
Response rate to teaching evaluation	0.294*	0.303*					
	(0.164)	(0.164)					
Course leader		0.150***	0.137***	0.063	0.065	0.151***	
		(0.043)	(0.037)	(0.055)	(0.055)	(0.039)	
Course leader × female				0.183**	0.184**		
				(0.083)	(0.083)		
Age						0.043*	
						(0.024)	
Age squared						-0.000*	
						(0.000)	
Observations	688	688	688	688	688	599	599
Course fixed effects	yes [95]	yes [95]	no	no	no	no	no
Year fixed effects	yes [4]	yes [4]	no	no	no	no	no
Course-year fixed effects	no	no	yes [272]	yes [272]	yes [272]	yes [261]	yes [261]

Table 4: Results controlling for observable course and teacher characteristics. Standard errors are clustered at the course level whenever we employ course fixed effects and at the course-year level whenever we employ course-year fixed effects. */ **/ *** $p < 0.10/0.05/0.01$, respectively. The average number of teachers per course is 2.53.

In sum, we find a negative effect of teacher gender on student evaluations of teaching once we properly control for course and teacher characteristics. The gender effect explains between 20.55 percent and 27.77 percent of the sample standard deviation across models. The variability of the estimates is small and robust to the conclusion of co-variates.

5.3 GENDER AND TEACHER QUALITY

Until now we have not controlled for teacher quality *per se* – an omitted variable, which may be correlated with gender and SETs. ISS has a strong commitment to research in teaching. Therefore, we consider research output a fair proxy of teacher quality as it reflects whether the teacher and researcher is up to date with the literature in her/his field. As a school of development studies, ISS relies on an integrated system of publication ratings and output valuation by the European Association of Development Institutes (EADI). The ranking system has been approved in 2006 and is used by EADI as performance valuation at the European level. The EADI ranking covers eight major scientific 'domains.'[7] Each type of publication gets credits depending on the impact factor of the book or journal, the refereeing procedure, and the number of authors. Publications are rated into five main categories from A to E. The A publications are within the top 33 percent of their domain according to the Thomson Reuter's ISI Web of Science.[8] For co-authored A publications researchers obtain four points, co-authored B publications are worth 3 points.

We first employ the sum of all EADI research points a researcher has gained with her/his publications in the full year prior to the evaluated course. Thus, for the academic year 2010/11 we employ the research points of the year 2010. When including the EADI research points in the main specification, we find a small positive correlation with the teaching evaluation (Table 5, Column 1). Since the descriptive statistics have shown that men tend to have a higher research output on average, it is important to control for the EADI research points to see whether the gender results are sensitive to such a specification. However, the link between overall research output and teaching quality is not statistically significant, while the link between gender and

[7] The eight domains are: (i) Anthropology and ethnic studies (and arts), (ii) area studies, planning and development, (iii) economics and management, (iv) geography, demography, environmental and urban studies (and natural and technical sciences), (v) political science, international relations, (international) law, public administration, history, (vi) sociology, social issues, (social) psychology, gender studies, (vii) psychology and health system studies (and medical sciences), and (viii) education and communication research.

[8] Further information on the EADI ranking system can be found in the CERES Research Valuation (2007). For the Thomson Reuter's ISI Web of Science we refer the reader to http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/web-of-science.html [Last access: February 2, 2016].

student evaluations of teaching remains unchanged. The coefficient estimate associated with gender is equal to -0.130.

Dependent variable: Teaching score	(1)	(2)	(3)	(4)
Female teacher	-0.130**	-0.134**	-0.157**	-0.061*
	(0.055)	(0.056)	(0.062)	(0.033)
Non-Caucasian teacher	0.051	0.039	0.057	0.045
	(0.046)	(0.045)	(0.046)	(0.033)
Course leader	0.145***	0.148***		0.083***
	(0.045)	(0.045)		(0.030)
EADI research points	0.002			
	(0.003)			
Number of A publications		0.028*		
		(0.017)		
Number of B publications		0.003		
		(0.017)		
Number of C publications		-0.013		
		(0.013)		
Average score per teacher				0.817***
				(0.065)
Observations	499	499	499	688
Course-year fixed effects	yes [249]	yes [249]	yes [249]	yes [272]

Table 5: Results controlling for research output of the teacher as measure of quality. Standard errors are clustered at the course-year level. */ **/ *** p < 0.10/0.05/0.01, respectively. The average number of teachers per course is 2.52.

Research quality rather than overall output may matter. We now separate research output into the number of A, B and C publications. We expect that teachers who are active in producing high-quality research are better teachers. Results are presented in Table 5, Column 2. A publications are positively and significantly associated with good student evaluations of teaching. B publications are similarly positively related and C publications have a negative sign. Yet the effect of B and C publications is imprecisely estimated. It does not seem to be research output per se but high quality research that feeds into more effective teaching; although the coefficient associated with high quality research is fairly small, namely 0.028. More importantly, the coefficient associated with gender remains unaffected by the inclusion of these covariates. To put effects in perspective: a female teacher would need almost 5 A publications

in the year prior to the evaluation to offset the direct negative gender effect in teaching evaluations.

Further note that we do not have information about publications for all teaching staff. Therefore, we had to carry out the analysis on a sub-sample including 499 teaching evaluations for 53 teachers. To verify the robustness of the initial results for the sub-sample we also estimate the basic model with course-year fixed effects and only the gender and ethnicity indicators as controls (Table 5, Column 3). The subsample results are identical to the results for the full sample (Table 3, Column 4). In other words, data on research quality is likely to miss at random from the sample.

5.4 GENDERED TRAITS

In our results we tried to disentangle teacher quality from the teaching scores by controlling for the research output of the teachers. However, even if the research profiles of the teachers perfectly capture the quality of their teaching in class, we cannot rule out that female and male teachers differ systematically in their personality types and traits and ultimately in the quality of teaching that they provide. Consequently, our point estimates might capture gendered traits instead of gender *per se*. One might argue that women are systematically less extrovert than men resulting in male teachers being the better 'entertainers' in front of class. If the personality traits are captured by the student evaluations of teaching, we should not be able to find any gender differences in teaching evaluations once we control for the average score a teacher obtains over time and across courses. In Table 5, Column 4, we present exactly this specification. Despite controlling for the average teaching score, the gender bias remains but is smaller in magnitude. Even in that specification, gender bias explains as much as 13.77 percent of the standard deviation in teaching scores. Yet, these estimates have to be taken with a grain of salt given that average scores are endogenous. Nevertheless, the findings indicate that personality, which is considered as being reflected in a professor's average teaching score, does not rule out gender differences in the teaching evaluations. Based on these findings, we argue that personality types are similarly distributed across gender and one can plausibly attribute the negative gender effect to discrimination.

Further evidence for this line of reasoning comes from the experimental literature that has dismissed many gender 'myths'. Women have been found to be neither more nor less socially

oriented but simply more sensitive in accounting for social conditions compared to men (Croson and Gneezy, 2009). When confronted with ethical dilemmas, at least on paper, women do not appear to have stronger ethical beliefs (Loo, 2003).[9] Even the widely held believe that women are more emotional has been dismantled (Feldman Barrett et al., 1998).

Dependent variable: Teaching score	(1)	(2)	(3)	(4)
Female teacher	-0.108**	-0.075**	-0.078**	-0.137**
	(0.047)	(0.034)	(0.035)	(0.059)
Non-Caucasian teacher	0.047	0.041	0.039	0.044
	(0.040)	(0.027)	(0.027)	(0.040)
Course leader	0.122***	0.134***	0.126***	0.126***
	(0.038)	(0.030)	(0.030)	(0.038)
Agrarian, food & environmental studies × female				-0.211
				(0.288)
Economics of development × female				-0.070
				(0.127)
Governance & development policy × female				-0.047
				(0.125)
Social justice perspectives × female				0.366***
				(0.128)
Social policy for development × female				0.138
				(0.120)
Research techniques courses × female				0.025
				(0.191)
Observations	651	652	615	688
Course-year fixed effects	yes [260]	yes [270]	yes [257]	yes [272]
Trimmed	5% from above	5% from below	5% from above & below	no

Table 6: Results trimming the sample from below and above and heterogeneity analysis across major core courses. Standard errors are clustered at the course-year level. */ **/ *** p < 0.10/0.05/0.01, respectively.

[9] Moreover, from experiments among Ugandan police officers we know that even in sensitive work environments women and men apply similar judgments (Wagner et al., 2016). Concerning corruption and gender, laboratory evidence suggests that women are not intrinsically more honest, but more opportunistic when they have the chance to break an implicitly corrupt contract. This results in lower corruption in mixed gender teams (Frank et al., 2011). The accumulated evidence on women's willingness to engage in corrupt behavior suggests that contextual factors matter rather than gender per se (Esarey, and Chirillo, 2013; Alatas et al., 2009; Alhassan-Alolo, 2007; Schulze and Frank, 2003; Sung, 2003).

5.5 CHAMPION TEACHERS

Our entire results might be driven by champions, i.e. teachers who tend to score very high on SETs and happen to be men. Similarly, we could have some female teachers in the sample who always score very low and drive our results. Therefore, we carried out further robustness tests trimming the sample from below and above. The results are presented in Table 6. Across specifications we control for gender, ethnicity and course leadership as the latter has proven to be an important factors explaining SET scores.[10]

In Column 1 of Table 6 we show that the negative and significant gender effect remains when we trim the top 5 percent of the observations. The coefficient is in line with the one from the main specification (Table 3, Column 4). Similarly, when we trim the bottom we obtain a negative and statistically significant gender effect (Table 6, Column 2). The coefficient does become smaller but explains still as much as 16.93 percent of the overall sample standard deviation. Trimming 5 percent of the observations from above and another 5 percent of the observations from below similarly does not render the negative gender effect insignificant (Table 6, Column 3). Across specifications the positive effect associated with course leadership remains. Clearly, our results are not driven by outliers.

5.6 PLACEBO TESTS

To further assess the robustness of our results we compare the individual teacher specific evaluations with the overall assessment of the course. If this overall evaluation is not sensitive to the gender composition of the teaching team, while the individual evaluations are, this can be interpreted as another piece of evidence of bias against women. This placebo analysis also checks if our identification strategy is driven by a special sub-sample of classes. If the likelihood that a male teacher is teaching alone, or only with other males is correlated with overall course ratings, then we are identifying the gender effect on individuals' evaluations of a special sub-sample with reduced external validity.

[10] The results are not driven by the inclusion of the variable 'course leader'.

Dependent variable	Overall perception of subject matter taught in the course	Value for professional development	Workload of the course	Time spent studying for the course
	(1)	(2)	(3)	(4)
Share of female teachers	-0.089	0.052	-0.176	0.065
	(0.160)	(0.114)	(0.142)	(0.051)
Share of Non-Caucasian teachers	-0.133	0.091	-0.087	-0.017
	(0.100)	(0.088)	(0.116)	(0.060)
Number of participants (log)	0.083	-0.052	0.062	0.014
	(0.089)	(0.067)	(0.124)	(0.038)
Response rate to teaching evaluation	-0.017	0.440***	0.527***	-0.035
	(0.187)	(0.159)	(0.169)	(0.077)
Control for change in question wording	0.439***			
	(0.131)			
Observations	270	236	256	269
Course fixed effects	yes [93]	yes [89]	yes [93]	yes [93]
Year fixed effects	yes [4]	yes [4]	yes [4]	yes [4]

Table 7: Results for the evaluation of general course aspects and workload. Standard errors are clustered at the course level. */ **/ *** p < 0.10/0.05/0.01, respectively.

In Column 1 of Table 7 we present the results associated with the evaluation criterion "Overall perception of the subject matter taught in the course." While the point estimates associated with the share of female or the share of Non-Caucasian teachers are negative, they are not statistically significant and we find no evidence that course contents delivered by mixed female, and ethnic minority teaching teams are evaluated differently.[11] The assessment of course-related criteria was less comprehensive in some years and courses compared to the individual teacher-level evaluations. Therefore, we do not have full information for all 272 courses on all outcomes.

We further assessed whether students consider a course valuable for their personal development. Again, we find that the valuations given by the students are unaffected by the composition of the teaching team (Table 7, Column 2).[12]

[11] Note that in this specification we control for a change in question wording that occurred in Term 2 of the academic year 2014/15.

[12] This question was only posed in 89 courses. Similarly questions regarding the workload and the time spent studying for the course were not asked for all courses. Compare the bottom of Table 7 for detailed information about the number of observations.

Last but not least we studied the feedback received from the students about the workload of the course and the time spent studying for the course relative to other courses (Table 7, Columns 3 and 4). Neither the perceived workload nor the time spent on the course are influenced by the composition of the teaching team. The lack of a correlation between general course characteristics and perceived workload with gender and ethnicity of the teaching team is a plausible cross-validation check for our main gender result.

5.7 COURSE HETEROGENEITY

Is there heterogeneity in the gender effect across the types of study programs? In particular, one out of the five majors taught –the social justice major– focuses on gender studies and it is reasonable to expect gender effects to be different in this program. One caveat is of course that it is impossible to disentangle self-selection of gender-sensitive students from the major content itself. But we can set up a reasonably rigorous empirical specification controlling for course-year fixed effects while including an interaction term between gender and the core courses of each of the five majors. We also include an interaction between gender and the more technical, research techniques courses. We treat the elective courses and the overarching courses as excluded category.

The heterogeneity results by major are presented in Column 4 of Table 6. The overall, gender bias remains. However, SETs from the core courses of the major in social justice reverse the gender effect. The overall gender effect is -0.137, the gender effect for the core courses in social justice perspectives is +0.366. The difference is statistically significant with a *p*-value of 0.042 suggesting that female teachers are better evaluated in a gender-sensitive context or by gender-sensitive students.

6 DISCUSSION AND CONCLUSION

Assessing the teaching performance of teachers in higher education is as important as it is challenging. Most university teachers, in particular young scholars, have to allocate their time between courses, research, management tasks and grant writing. Against this background it is important to study the explanatory power and causal determinants of teaching evaluations. Teaching evaluations –as in the context of this study– are often used in promotion and hiring

decisions. Our results suggest that fair evaluations need to net out gender discrimination, the role of the teacher in the course, class size and the research activities of the teacher. In line with previous findings (Boring et al., 2016; MacNell et al., 2014; Basow and Silberg, 1987), our results indicate that teacher performance is assessed more critically for female teachers, even after controlling for a plethora of cofounding factors.

More generally our findings that are based on a novel identification strategy exploiting within course variation also complement previous work on gender bias in the academic work environment as a whole. Carrell et al. (2010) demonstrate that in science, technology, engineering, and mathematics the gender of teachers has a considerable impact on female students' performance but not on male students. When high-performing female students are matched with female professors, the gender gap in these majors vanishes. Similar evidence for the role of respondent gender is provided by evaluations collected in teaching hospitals: Male resident doctors of 20 hospitals in the Netherlands evaluated their clinical teachers better than female residents (Arah et al., 2012). In contrast, van der Leeuw et al. (2013) present evidence that female medical instructors tend to give a more detailed narrative feedback both in terms of positive comments as well as suggestions for improvement. Similarly, based on administrative data from a university in California, USA it was found that the ethnicity of teaching assistants has a positive and significant impact on course grades when the students and the teaching assistant share the same racial background (Lusher et al., 2015). Last but not least, our research not only contributes to the perception of women in the university environment, but also to the broader literature on gender discrimination in the labor market.[13]

Before concluding we would like to point out caveats of this paper. Most importantly, due to privacy concerns we have no auxiliary background information about the respondents, i.e. the students. Previous research has shown that the gender of the student influences evaluations (Boring et al., 2016). And this student gender effect may interact with the gender of the teacher. If, in addition, male and female students as well as male and female teachers select into certain

[13] Labor market studies over and again point to a negative gender effect on wages (see for instance, Azier, 2010; Weichselbaumer and Winter-Ebmer, 2005; Blau and Kahn, 2000). A wide range of explanations have been put forward to construe this pay gap, ranging from gendered occupational choices (Turner and Bowen, 1999; Petersen and Morgan, 1995) to more favorable promotion opportunities for men (Arulampalam et al., 2007) to name just two. But stereotyping and social constructs also contribute to the discrimination of women (Andreoni and Petrie, 2008; Fiske, 2000; Oakes et al., 1994) and might ultimately feed into the construction of observed gender gaps.

classes based on gender lines, then part of the documented gender effect might be driven by gendered sorting of students. If female students tend to select into classes taught by female professors, then our results are downward biased. Finally, it would be interesting to evaluate the causal impact of raising awareness for gender-bias on student evaluations.

Our paper has several policy implications: First, our estimates could be used to adjust evaluations accordingly given that student evaluations of teaching are still employed as the predominant indicator to assess teaching effectiveness (Becker et al., 2012; McPherson, 2010; Davies et al., 2007; Emery et al., 2003). Yet, we acknowledge that it is questionable whether a fair metric, which is widely supported, can be found. Second, narrative, qualitative feedback could also be taken into account next to quantitative measures. Cut-off points for excellence in teaching, as it is practice in our study setting, are arbitrary and need to be complemented with qualitative feedback in order to get a holistic picture about teacher performance in class. Third, evaluations could not only be used for in-class teaching but for other teacher activities such as student supervision as well. More introvert and possibly female teachers might be less well perceived in class, but perform rather well in one-to-one interactions (Arah et al., 2012). Finally, alternative ways of evaluating the performance of teachers could be used to minimize biases. One possibility may be peer review by both female and male colleagues (Baldwin and Blattner, 2003; Sproule, 2002).

REFERENCES

Abrami, P. C and S. d'Apollonia. 1991. Multidimensional students' evaluations of teaching effectiveness -- Generalizability of "*N* = 1" research: Comment on Marsh (1991). Journal of Educational Psychology 83(3): 411–415.

Abrami, Philip C., Les Leventhal, Raymond P. Perry, and Lawrence J. Breen. 1976. Course Evaluation: How?. Journal of Educational Psychology 68(3): 300–304.

Alatas, Vivi, Lisa Cameron, and Ananish Chaudhuri. 2009. Gender, Culture, and Corruption: Insights from an Experimental Analysis. Southern Economic Journal 75(3): 663–80.

Alhassan-Alolo, Namawu. 2007. Gender and Corruption: Testing the New Consensus. Public Administration and Development 237(3): 227–37.

Andreoni, James and Ragan Petrie. 2008. Beauty, gender and stereotypes: Evidence from laboratory experiments. Journal of Economic Psychology 29(1): 73–93.

Arah, Onyebuchi A, Maas J Heineman and Kiki M J M H Lombarts. 2012. Factors influencing residents' evaluations of clinical faculty member teaching qualities and role model status. Medical Education: 46: 381–389.

Arulampalam, Wiji, Alison L. Booth and Mark L. Bryan. 2007. Is There a Glass Ceiling over Europe? Exploring the Gender Pay Gap across the Wage Distribution. Industrial and Labor Relations Review 60(2): 163–186.

Azier, Anna. 2010. The Gender Wage Gap and Domestic Violence. American Economic Review 100(4): 1847–1859.

Bachen, Christine M., Moira M. McLoughlin and Sara S. Garcia. 1999. Assessing the role of gender in college students' evaluations of faculty. Communication Education 48(3): 193–210.

Baldwin, Tamara and Nancy Blattner. 2003. Guarding Against Potential Bias in Student Evaluations: What Every Faculty Member Needs to Know. College Teaching 51(1): 27–32.

Basow, S. A. and N. T. Silberg. 1987. Student evaluations and college professors: Are female and male professors rated differently? Journal of Educational Psychology 79: 308-314.

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

Becker, W. E. and M. Watts. 1999. How departments of economics should evaluate teaching. American Economic Review (Papers and Proceedings) 89: 344–349.

Beleche, Trinidad, David Fairris and Mindy Marks. 2012. Do course evaluations truly reflect student learning? Evidence from an objectively graded post-test. Economics of Education Review 31(5): 709–719.

Blau, Francine D. and Lawrence M. Kahn. 2000. Gender Differences in Pay. Journal of Economic Perspectives 14(4):75–99.

Boring, Anne, Kellie Ottoboni and Philip B. Stark. 2016. Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness. Working Paper.

Braga, Michela, Marco Paccagnella and Michele Pellizzari. 2014. Evaluating students' evaluations of professors. Economics of Education Review 41: 71–88.

Cadwell, J. and J. Jenkins. 1985. Effects of semantic similarity of items on student ratings of instructors. Journal of Educational Psychology 77: 383–393.

Carrell, Scott E., Marianne E. Page and James E. West. 2010. Sex and Science: How Professor Gender Perpetuates the Gender Gap. The Quarterly Journal of Economics 125(3): 1101–1144

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

Cashin, William E. 1990. Students Do Rate Different Academic Fields Differently. In M. Theall and J. Franklin (Eds.), Student ratings of instruction: Issues for improving practice, New Directions for Teaching and Learning, No. 43. San Francisco: Jossey-Bass, pp. 113–21.

CERES. 2007. CERES Research Valuation. The CERES (and EADI) system of valuation of research output. Fact Sheet.

Crockett, Emily. 2015, November 12. Even famous women economists get no respect. This online quiz shows why. VOX. Retrieved from http://www.vox.com/2015/11/12/9724796/female-economists-respect-quiz.

Croson, Rachel and Uri Gneezy. 2009. Gender Differences in Preferences. Journal of Economic Literature 47(2): 448–74.

Davies, Martin, Joe Hirschberg, Jenny N. Lye, Carol Georgina Johnston and Ian M. McDonald. 2007. Systematic Influences on Teaching Evaluations: The Case for Caution. Australian Economic Papers 46(1): 18–38.

Esarey, Justin and Gina Chirillo. 2013. 'Fairer Sex' or Purity Myth? Corruption, Gender, and Institutional Context. Politics & Gender 9: 361–389.

Emery, C.R., T.R. Kramer and R.G. Tian. 2003. Return to academic standards: A critique of student evaluations of teaching effectiveness. Quality Assurance in Education 11(1): 37–46.

Ewing, Andrew M. 2012. Estimating the impact of relative expected grade on student evaluations of teachers. Economics of Education Review 31(1): 141–154.

Feldman Barrett, Lisa, Lucy Robin, Paula R. Pietromonaco and Kristen M. Eyssell. 1998. Are Women the 'More Emotional' Sex? Evidence From Emotional Experiences in Social Context. Cognition and Emotion 12(4): 555–578.

Frank, Björn, Johann Graf Lambsdorff and Frédéric Boehm. 2011. Gender and Corruption: Lessons from Laboratory Corruption Experiments. European Journal of Development Research 23: 59–71.

Fiske, Susan T. 2000. Stereotyping, prejudice, and discrimination at the seam between the centuries: evolution, culture, mind, and brain. European Journal of Social Psychology 29: 299–322.

Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology* 74(6): 1464–1480.

Guthrie, E. R. 1954. The Evaluation of Teaching: A Progress Report. Seattle: University of Washington.

Hamermesh, Daniel S. and Amy Parker. 2005. Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity. Economics of Education Review 24(4): 369–376.

Hay, Carol. 2016, January 25. Girlfriend, Mother, Professor? The New York Times Opinionator. Retrieved from http://opinionator.blogs.nytimes.com/2016/01/25/girlfriend-mother-professor/?_r=0. Last access: February 2, 2016.

Isely, Paul and Harinder Singh. 2005. Do Higher Grades Lead to Favorable Student Evaluations?. Journal of Economic Education 36(1): 29–42.

Johnson, V.E. 2003. Grade inflation: A crisis in college education. Springer-Verlag, New York, New York.

Kamenetz, Anaya. 2016, January 25. Why Female Professors Get Lower Ratings. National Public Radio (NPR) Education. Retrieved from http://www.npr.org/sections/ed/2016/01/25/463846130/why-women-professors-get-lower-ratings. Last access: February 2, 2016.

Krautmann, A.C. and W. Sander. 1999. Grades and student evaluations of teachers. Economics of Education Review 18: 59–63.

Langbein, Laura. 2008. Management by results: Student evaluation of faculty teaching and the mis-measurement of performance. Economics of Education Review 27(4): 417-428.

Loo, Robert. 2003. Are women more ethical than men? Findings from three independent studies. Women in Management Review 18(4): 169–181.

Lusher, Lester, Doug Campbell and Scott Carrell. 2015. TAs Like Me: Racial Interactions between Graduate Teaching Assistants and Undergraduates. National Bureau of Economic Research, Working Paper 21568.

MacNell, Lillian, Adam Driscoll and Andrea N. Hunt. 2014. What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. Innovative Higher Education.

Marsh, Herbert W. 2007. Do University Teachers Become More Effective With Experience? A Multilevel Growth Model of Students' Evaluations of Teaching Over 13 Years. Journal of Educational Psychology 99(4): 775–790.

Marsh, Herbert W. and Lawrence A. Roche. 2000. Effects of Grading Leniency and Low Workload on Students' Evaluations of Teaching: Popular Myth, Bias, Validity, or Innocent Bystanders?. Journal of Educational Psychology 92(1): 202–228.

Marsh, Herbert W. 1991. A multidimensional perspective on students' evaluations of teaching effectiveness: Reply to Abrami and D'Apollonia (1991). Journal of Educational Psychology 83(3): 416–421.

Marsh, Herbert W. 1987. Student Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research. International Journal of Educational Research 11: .253–388.

Marsh, Herbert W., and Mark A. Groves. 1987. Students' evaluations of teaching effectiveness and implicit theories: A critique of Cadwell and Jenkins (1985). Journal of Educational Psychology 79(4): 483–489.

Marsh, H. W. 1984. Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. Journal of Educational Psychology 76(5): 707-754.

McPherson, Michael A. 2006. Determinants of How Students Evaluate Teachers. Journal of Economic Education 37(1): 3–20.

Mo, Cecilia Hyunjung. 2014. The Consequences of Explicit and Implicit Gender Attitudes and Candidate Quality in the Calculations of Voters. Political Behavior 37(2): 357—395.

Oakes, Penelope J., S. Alexander Haslam, John C. Turner. 1994. Stereotyping and social reality. Malden: Blackwell Publishing Stereotyping and social reality.

Ongeri, Joseph D. 2009. Poor Student Evaluation of Teaching in Economics: A Critical Survey of the Literature.Australasian Journal of Economics Education 6(2): 1–24.

Petersen, Trond and Laurie A. Morgan. 1995. Separate and Unequal: Occupation-Establishment Sex Segregation and the Gender Wage Gap. American Journal of Sociology 101(2): 329–365.

Poropat, Arthur. 2014, November 18. Students don't know what's best for their own learning. The Conversation. Retrieved from http://theconversation.com/students-dont-know-whats-best-for-their-own-learning-33835. Last access: February 3, 2016.

Pounder, James S. 2007. Is student evaluation of teaching worthwhile?: An analytical framework for answering the question. Quality Assurance in Education 15(2): 178 – 191.

Schmidt, Ben. 2015. Gendered Language in Teaching Evaluations. Online Tool. Retrieved from http://benschmidt.org/profGender/#. Last access: February 25, 2016.

Schulze, Günther G. and Björn Frank. 2003. Deterrence versus Intrinsic Motivation: Experimental Evidence on the Determinants of Corruptibility. Economics of Governance 4(2): 143–60.

Seldin, P. 1993. The use and abuse of student ratings of professors. The Chronicle of Higher Education 39(46): A40.

Sproule, Robert. 2002. The underdetermination of instructor performance by data from the student evaluation of teaching. Economics of Education Review 21(3): 287–294.

Stark, Philip B. and Richard Freishtat. 2014. An Evaluation of Course Evaluations. Working Paper.

Sung, Hung-En. 2003. Fairer Sex or Fairer System? Gender and Corruption Revisited Social Forces 82(2): 703–723.

Turner, Sarah E. and William G. Bowen. 1999. Choice of Major: The Changing (Unchanging) Gender Gap. Industrial and Labor Relations Review. 52(2): 289–313.

van der Leeuw, Renée M., Karlijn Overeem, Onyebuchi A. Arah, Maas Jan Heineman and Kiki M.J.M.H. Lombarts. 2013. Frequency and Determinants of Residents' Narrative Feedback on the Teaching Performance of Faculty: Narratives in Numbers. Academic Medicine 88(9): 1324–1331.

Wagner, Natascha, Matthias Rieger and Arjun Bedi. 2016. Corruption, Policing and Gender: Evidence from Survey Experiments in Uganda. ISS Working Paper No 615.

Walstad, William and Saunders, Phillip. 1998. Using Student and Faculty Evaluations of Teaching To Improve Economics Instruction. In William Walstad and Phillip Saunders (Eds.), Teaching undergraduate economics: A handbook for instructors. Boston: Irwin McGraw-Hill, pp. 337–55.

Weichselbaumer, Doris and Rudolf Winter-Ebmer. 2005. A Meta-Analysis of the International Gender Wage Gap. Journal of Economic Surveys 19(3): 479–511.

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