Gender, ethnicity and teaching evaluations: Evidence from mixed teaching teams

Natascha Wagner*, Matthias Rieger, Katherine Voorvelt

International Institute of Social Studies, Erasmus University Rotterdam, Kortenaerkade 12, 2518AX, The Hague, Netherlands

ARTICLE INFO

ABSTRACT

This paper studies the effect of teacher gender and ethnicity on student evaluations of teaching at university. We analyze a unique data-set featuring mixed teaching teams and a diverse, multicultural, multi-ethnic group of students and teachers. Blended co-teaching allows us to study the link between student evaluations of teaching and teacher gender as well as ethnicity exploiting *within* course variation in a panel data 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 compared to men. The effect is robust to a host of co-variates such as course leadership, teacher experience and research quality, as well as an alternative teacher fixed effect specification. There is no evidence of a corresponding ethnicity effect. Our results are suggestive of a gender bias against female teachers and indicate that the use of teaching evaluations in hiring and promotion decisions may put female lectures at a disadvantage.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

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

Existing research does suggest that the resulting average evaluations are reliable and stable, but they are to a large extent a function of teacher characteristics and behavior rather than course content and quality *per se* (Marsh, 1987; Pounder, 2007). Whether student evaluations of teaching are related to course grades and workload is contested.[1] At the same time, only a modest positive association between teaching evaluations and student learning has been found (see Beleche, Fairris, & Marks, 2012).[2]

---

[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 & West, 2010; Weinberg, Fleisher & Hashimoto, 2009; Langbein, 2008; Isely & Singh, 2005; Johnson, 2003; Krautmann & Sander, 1999).

[2] Braga, Paccagnella and Pellizzari (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% of the variation in

---

* Corresponding author.: + 31704260574.
E-mail addresses: wagner@iss.nl (N. Wagner), rieger@iss.nl (M. Rieger), voorvelt@oos.eur.nl (K. Voorvelt).

Apart from these caveats, the use of average SET scores ignores issues related to response rates and response variability (Stark & 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, McGhee, & Schwartz, 1998), which revealed implicit bias against women in positions of power (Crockett, 2015; Mo, 2014).

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, Ottoboni, & Stark, 2016; Seldin, 1993; Walstad & Saunders, 1998).

In this paper we focus on two important teacher traits and their potentially negative correlation with SETs: being female and being of Non-Caucasian ethnic background. 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. Their 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 (Basow & Silberg, 1987; Boring et al., 2016).

We propose a new identification strategy to assess the association between teacher traits and student evaluations of teaching. Average scores differ by subject and a naïve analysis where one combines all courses therefore cannot reveal gender-differences as suggested by Schmidt (2015).[3] We make use of a study setting where most lecturers 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 (Cashin, 1990; Ongeri, 2009; Schmidt, 2015).

We document significantly lower scores in teaching evaluations for women compared to men, but only once we control for course unobservables. In other words, the documented associations insinuate that teacher evaluations are not gender blind, and gender effects explain roughly one fourth of the sample standard deviation in SETs. Our findings suggest that women are 11 percentage points less likely to attain the teaching evaluation cut-off for promotion to associate professor compared to men. Our results are also robust to netting out teacher unobservables (such as ability or personality) in an alternative panel model specification. More specifically, we run teacher fixed effect models separately for men and women. This allows us to estimate the impact of co-teaching with the opposite, as well as same gender relative to teaching a course alone. Women obtain considerably lower teacher evaluations when teaching with men compared to teaching alone or with other women.

The negative female teacher effect is also important in magnitude compared to other significant correlates of SETs such as research productivity 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, Driscoll, and Hunt (2014). The crux of this 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 female teacher effect is reversed in the major for gender studies and social justice. Finally, we cross-validate our main findings by looking at the effect of teaching team composition on overall evaluations of courses. While gender seems to matter for individual evaluations of teachers, the share of female teachers and the composition of teaching teams (female-only and mixed gender teams versus male-only teams) have no systematic effect on how students perceive the course in general. If at all there is a weak, positive association between female-only and mixed gender teams and the perceived overall course quality. We interpret these patterns as additional, suggestive evidence of bias against female teachers.

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

---

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 & 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, Leventhal, Perry & Breen, 1976).

[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.

[4] Related, Bachen, McLoughlin and Garcia (1999) argue 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.