Bias in student evaluations of teaching

DOMINIC VECONI

PENN STATE UNIVERSITY DEPARTMENT OF MATHEMATICS

OCTOBER 31, 2017

Background

- Student evaluations of teaching (SET) are often used in hiring, promoting or firing faculty in different universities.
- Not a clear (or even positive) correlation between actual teaching effectiveness and SET results (Mengel et al. (2016), Boring et al. (2016))
- One would expect that an effective teacher has students who perform better in follow-up courses, for example, and that this would positively correlate with good SETs. This is often not the case (in fact sometimes this correlation is negative).

What do good SET results correlate with?

- **Grade expectation**: Students who believe they are getting a good grade tend to give higher SETs.
- **Gender bias**: Male students tend to give better SETs to male instructors, despite little to no difference between academic performance of male students of male v. female instructors (Mengel et al., Boring et al.)

Defining bias in SETs

- Centra and Gaubatz: Bias (in groups of students) occurs when "a teacher or course characteristic affects teacher evaluations, either positively or negatively, but is unrelated to the criteria of good teaching, such as increased student learning."
- Mengel, Sauermann, and Zölitz: Gender biases (in individual student SETs) are "gender differences in evaluations which cannot be explained via grades or effort…."

Methods: Boring, Ottoboni, Stark (2016):

- Examined whether SETs correlate to actual teaching effectiveness, or bias of some kind
  - Main bias came from grade expectation, and instructor gender.
  - Also stratified data based on discipline and gender of the students.
- Nonparametric permutation tests: Avoid contrived assumptions about parametric generative models for the data (which linear regression, *t*-tests, and ANOVA typically require)
- Null hypothesis: Certain instructor characteristics (i.e. gender) are arbitrary labels, and the correlation is comparable to the correlation attained from a random assignment of labels "male" and "female."

Methods: Boring, Ottoboni, Stark:

- Data comes from two main experimental data sets:
  - French natural experiment
  - US randomized experiment
- French natural experiment:
  - 23,001 SET of 379 instructors (34% women) from 4,423 students (57% women)
  - Students randomly assigned to different sections, with different instructors for different courses
  - Microeconomics, macroeconomics, history, political institutions, political science, sociology

Methods: Mengel, Sauermann, Zölitz:

- Examined 19,962 SET results for 735 teachers at the School of Business and Economics (SBE) of Maastricht University, Netherlands, across academic years 2009/2010 – 2012/2013
- Students randomly assigned to different sections
- 35% of instructors and 38% of students were women
- Mostly business or economics students, mixture of bachelor, master and doctoral students

Methods: Mengel, Sauermann, Zölitz:

- Developed conceptual framework to describe different factors that influence a student's end-of-semester SET.
- Three qualitative variables affect SET results: GRADE (final performance), EFFORT (hours put in outside of class), and EXPERIENCE (other factors).
- Null hypotheses:
  A. There are no gender differences between male and female student evaluations for male or female instructors
  B. Neither male nor female students differ in their evaluations of male or female instructors (but male student evaluations and female student evaluations may differ generally)
  C. Female students do not evaluate male or female instructors differently based on their performance
  D. Male students do not evaluate male or female instructors differently based on their performance

Results: French (Boring et al.)

- Correlation between SET and objective effectiveness measure (i.e. performance on anonymously graded final exams) is weak and often not statistically significant.
- Correlation between SET and instructor gender is significant overall, but less significant within individual disciplines.
- Correlation between instructor gender and final grade is insignificant, but slightly favors female instructors.

Subject	SET/final grade av. cor.	SET/instructor gender	Final grade/instructor gender
Overall	$0.04\ (p = 0.09)$	$0.09\ (p = 0.00)$	$-0.06\ (p = 0.07)$
History	$0.16\ (p = 0.01)$	$0.11\ (p = 0.08)$	$-0.08\ (p = 0.22)$
Political institutions	N/A	$0.11\ (p = 0.10)$	NA
Macroeconomics	$0.06\ (p = 0.19)$	$0.10\ (p = 0.16)$	$-0.06\ (p = 0.37)$
Microeconomics	$-0.01\ (p = 0.55)$	$0.09\ (p = 0.16)$	$-0.06\ (p = 0.37)$
Political science	$-0.03\ (p = 0.62)$	$0.04\ (p = 0.63)$	$-0.03\ (p = 0.70)$
Sociology	$-0.02\ (p = 0.61)$	$0.08\ (p = 0.34)$	$-0.05\ (p = 0.55)$

Results: French (Boring et al.)

- Gender concordance between instructor and student correlates with SET scores (much better predictor for male students than female students)
- However, correlation between gender concordance and final exam grade is weak to insignificant; in fact, male history students peform *worse* on the final exam for male instructors significantly, despite giving male instructors better SET results.

Subject	SET/gender concordance		Final grade/gender concordance	
	Male student	Female student	Male student	Female student
Overall	$0.15\ (p = 0.00)$	$0.05\ (p = 0.09)$	$-0.1\ (p = 0.75)$	$0.06\ (p = 0.07)$
History	$0.17\ (p = 0.01)$	$-0.03\ (p = 0.60)$	$-0.15\ (p = 0.03)$	$-0.02\ (p = 0.74)$
Political institutions	$0.12\ (p = 0.08)$	$-0.11\ (p = 0.12)$	N/A	N/A
Macroeconomics	$0.14\ (p = 0.04)$	$-0.05\ (p = 0.49)$	$0.04\ (p = 0.60)$	$0.11\ (p = 0.10)$
Microeconomics	$0.18\ (p = 0.01)$	$-0.00\ (p = 0.97)$	$0.02\ (p = 0.80)$	$0.07\ (p = 0.29)$
Political science	$0.17\ (p = 0.06)$	$0.04\ (p = 0.64)$	$0.08\ (p = 0.37)$	$0.11\ (p = 0.23)$
Sociology	$012\ (p = 0.16)$	$-0.03\ (p = 0.76)$	$0.01\ (p = 0.94)$	$0.06\ (p = 0.47)$

Results: Dutch (Mengel et al.)

- Bias in evaluations (numbers are multiple of standard deviation differences):

Student gender	Teacher-related (1)	Group-related (2)	Material-related (3)	Course-related (4)	Hours spent (5)	Final grade (6)
Men	-0.2070**	-0.0576*	-0.0569*	-0.0760**	0.0459	0.0115
Women	-0.0769*	-0.00932	-0.0317	-0.0240	-0.0465	0.0395

- $^{**}p < 0.01$, $^{*}p < 0.05$
- Men rate female instructors 20.7% SD worse than male instructors.
  - Standard deviation of evaluation items from (1): 0.93, i.e. 0.21 point difference on 5-point Likert scale
- Columns (2), (3), (4) based on evaluation questions unrelated to instructor, but women are still rated worse as instructors
- Any system that ranks instructors (0 being worst and 1 being best) would hypothetically, using this data, translate to 0.37 lower rank for female instructors on average.

Results: Dutch (Mengel et al.)

- Bias in evaluations (numbers are multiple of standard deviation differences):

Student gender	Teacher-related (1)	Group-related (2)	Material-related (3)	Course-related (4)	Hours spent (5)	Final grade (6)
Men	-0.2070**	-0.0576*	-0.0569*	-0.0760**	0.0459	0.0115
Women	-0.0769*	-0.00932	-0.0317	-0.0240	-0.0465	0.0395

- $^{**}p < 0.01$, $^{*}p < 0.05$
- Example of rank-based outcome: SBE Teaching Awards (3 categories):
  1. Student teachers (40% women in category, 15% women nominees)
  2. Undergraduate teaching (38% women in category, 26% women nominees)
  3. Graduate teaching (32% women in category, 27% nominees)
- Instructor gender has no significant impact on EFFORT or GRADE variables. Therefore, SET differences must come from EXPERIENCE variable.

Age differences (Mengel et al.)

Male students rank female instructors who are masters and PhD students on average 27.11% and 28.01% of a standard deviation lower than male instructors.

Female students rank female masters students 31.49% lower than male instructors, but rank female lecturers and professors 13.78% and 27.25% *higher* than male lecturers and professors.

Table 6: Estimates for male students (β_1; Panel 1) and female students ($\beta_1+\beta_3$; Panel 2) depending on teacher and student seniority.

	→ Increasing Seniority Teacher →				
	Student	PhD student	Lecturer	Professor	Overall
	Panel 1: Male Students ($\widehat{\beta}_1$)				
1st year Bachelor	-0.2304	-0.3488**	-0.1083**	0.1982	-0.0941
2nd year Bachelor and higher	-0.2744	0.1528	-0.0304	0.1436	-0.1970***
Master	-0.5068**	-0.6346***	0.2044	-0.0178	-0.2645***
Overall	-0.2711***	-0.2801***	-0.0425	0.1029	-0.1839***
	Panel 2: Female Students ($\widehat{\beta}_1 + \widehat{\beta}_3$)				
1st year Bachelor	-0.2822**	-0.2570	-0.0162	0.5680***	-0.0347
2nd year Bachelor and higher	-0.3399***	0.2309**	0.2207**	0.3930**	0.0046
Master	-0.5130***	-0.4584	0.3383*	0.1013	-0.1248*
Overall	-0.3149***	-0.1341	0.1378*	0.2725**	-0.0391
	Panel 3: Number of observations				
1st year Bachelor	1523	1218	1600	303	4644
2nd year Bachelor and higher	1933	1878	2527	1497	7835
Master	448	1707	1365	2244	5764
Overall	3904	4803	5492	4044	18243

Note: *** p<0.01, ** p<0.05, * p<0.1. Dependent variable: Teacher evaluation. All estimates are based on regressions which include course fixed effects, parallel course fixed effects for the courses taken at the same time, and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level in parentheses.

Can age difference be explained?

- Does academic seniority convey a sense of authority that junior women lack?
  - May be interesting to look at differences between junior and senior male instructors, and junior and senior female instructors.
- Are women senior faculty better instructors? Is there a stronger, more competitive "selection" effect for women than for men? (No statistically significant difference in GRADE or EFFORT in Mengel et al.)

	Student teacher	PhD student	Lecturer	Professor
Hours spent				
Male students	-0.0494	-0.5664	0.5975	0.4391
Female students	-0.174	0.162	-0.102	0.700
Grade received				
Male students	0.0131	0.0232	-0.1034	0.0842
Female students	-0.0599	0.0026	-0.0629	0.0210
Total observations	3,904	4,803	5,492	4,044

Teacher Added Value

- Teacher Added Value: A measure of effectiveness in education literature
  - Based on regression of students' GPA and grades in course the instructor is teaching
- Mengel et al.: Significantly lower SETs for women instructors than men in bottom three quartiles; suggests teacher quality has only weak effect on SET results
- $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Student gender:	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Male	-0.2296***	-0.2222***	-0.2941***	-0.0444
Female	-0.169*	-0.104	-0.195**	0.0274
Observations	4,984	4,864	4,962	5,152

Stereotype Threat

- Do the differences in SETs for women derive from negative stereotypes about of women in math?
  - If so, are students doubting the competence of their instructors? Or do female instructors lack confidence because of perceived negative stereotypes (stereotype threat)?
  - Latter is unlikely (Mengel et al.) ($^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$):

Student gender	Teacher evaluation		Hours spent		Grade	
	No math	Math	No math	Math	No math	Math
Male	-0.1723***	-0.3180***	0.0223	0.1347	0.0179	0.0306
Female	-0.0363	-0.277***	-0.0582	-0.0883	0.0600*	-0.0760
Observations	14,852	4,821	14,852	4,821	14,852	4,821

Bias in gender-incongruent areas?

- If there's SET bias against women as math instructors, what about bias against men in areas stereotypically dominated by women (e.g. education studies)?
- Mengel et al.: Bias effect size is comparable and in the same direction regardless of whether there are more male or female instructors for a particular course.
- $^{***}p < 0.01$, $^{*}p < 0.1$

Student gender	Majority male instructors	Majority female instructors
Male	-0.1793***	-0.2731***
Female	-0.0757*	-0.0749
Observations	14,300	5,662

Alternative outcomes and survey response rate (Mengel et al.)

- Still possible that male teachers perform better with respect to other learning outcomes that are harder to measure in exams
- But gender bias is much stronger among male students, so this would imply that male, but not female, instructors teach "towards" male students. Educational research is "only partially consistent" with this hypothesis
- Many studies (Altermatt et al. (1998), Jones and Dindia (2004), Halim and Ruble (2010)) find that both men and women treat male students preferentially. Suggests bias comes from preconceived stereotypes
- Female students and better-performing students more likely to respond to SBE's SET surveys.
- However, there is no significant correlation between teacher gender and response rate for male students, and having a female instructor leads to only a small increase in response rate for female students (not significant when controlling for grades/GPA).
- So alternative learning outcomes and survey response rates do not seem to drive disparities in ratings for female v. male instructors.

Other studies:

- Arbuckle and Williams (2003): 352 students ranked instructors they were told were men, women, young, and old. Young men got higher rankings than other three combinations.
- Race: Minority instructors get lower SET results than white instructors (Merritt (2008))
- Age, charisma, physical attractiveness in different studies as well.
- Classroom size, class time also affect the teacher's evaluations, even though instructor has no control over these factors.
- Overall, student satisfaction is a more significant contributor to teaching effectiveness on SET surveys (in both Boring et al. and Mengel et al., i.e. the EXPERIENCE factor in the Mengel et al. framework) than actual teaching effectiveness.

Possible negative effects:

- (Mengel et al.) SETs are often not corrected for possible gender bias or student gender composition.
- At SBE, found effect sizes were significant enough to damage chances of women to receive teaching awards, and perception among supervisors and colleagues.
- Could also have negative effect on female junior faculty's or female PhD students' confidence as instructors.

When can SETs be used?

- SETs have two common uses:
  - Informing instructors on how to improve their course
  - Influencing personnel decisions for departments
- The first one can be meaningful, but perhaps more as a way to get info from individual surveys on specific information for the instructor's class and their individual teaching style.
- Several studies recommend caution for using SETs in personnel decisions (Mengel et al.), or discontinuing their use entirely (Boring et al.), suggesting "the onus should be on universities that rely on SET for employment decisions to provide convincing affirmative evidence that such reliance does not have disparate impact on women, underrepresented minorities, or other protected groups". (Boring et al.)
- Since bias in SETs can come from several different places, and there is so much variance in the magnitude of the bias by subject, student gender, and evaluation item, there is little to no practical way to accurately and uniformly adjust for these biases.
- French and Dutch natural experiments have educational environments that closely resemble PSU. While this is not indicative of explicit bias at Penn State in our SRTEs, we should not think ourselves immune.

References

1. Boring A, Ottoboni K, Stark P. Student evaluations of teaching (mostly) do not measure teaching effectiveness. ScienceOpen Research 2016
2. Mengel F, Saureman J, Zölitz U. Gender bias in teaching evaluations. Institute of Labor Economics 2016; IZA DP No.11000
3. Centra J, Gaubatz N. Is there gender bias in student evaluations of teaching? J High Educ. 2000; 71(1):17-33
4. MacNell L, Driscoll A, Hunt AN. What's in a name? Exposing gender bias in student ratings of teaching. Innovat High Educ. 2015; 40(4):291-303
5. Altermatt E, Jovanovic J, Perry M. Bias or responsivity? Sex and achievement-level effects on teachers' classroom questioning practices. J Educ. Psych. 1998; 90(3):516-527
6. Halim M, Ruble D. Gender identity and stereotyping in early and middle childhood. Handbook of Gender Research in Psychology: Gender Research in General and Experimental Psychology 2010; 1:495-525
7. Jones S, Dindia K. A meta-analytic perspective on sex equity in the classroom. Rev. Educ. Research 2004; 74(4):443-471
8. Arbuckle J, Williams B. Students' perceptions of expresiveness: age and gender effects on teacher evaluations. Sex Roles. 2003; 49:507-516
9. Merritt D. Bias, the brain, and student evaluations of teaching. St. John's Law Rev. 2008; 81(1):235-288