Phi:male Coffee Lecture: Gender Bias in Teaching Evaluations

Sara Kališnik

December 19, 2022

BSc degree:
University of Ljubljana

BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 46, Number 2, April 2009, Pages 255–308
S 0273-0979(09)01249-X
Article electronically published on January 29, 2009

TOPOLOGY AND DATA

GUNNAR CARLSSON

1. Introduction

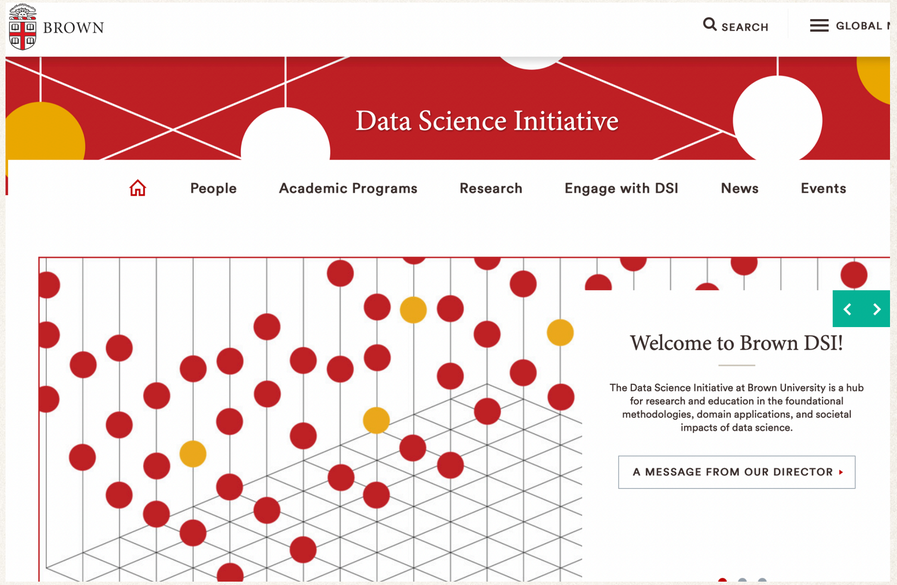
An important feature of modern science and engineering is that data of various kinds is being produced at an unprecedented rate. This is so in part because of new experimental methods, and in part because of the increase in the availability of high powered computing technology. It is also clear that the *nature* of the data we are obtaining is significantly different. For example, it is now often the case that we are given data in the form of very long vectors, where all but a few of the

Inspiration

PhD:
Stanford
University

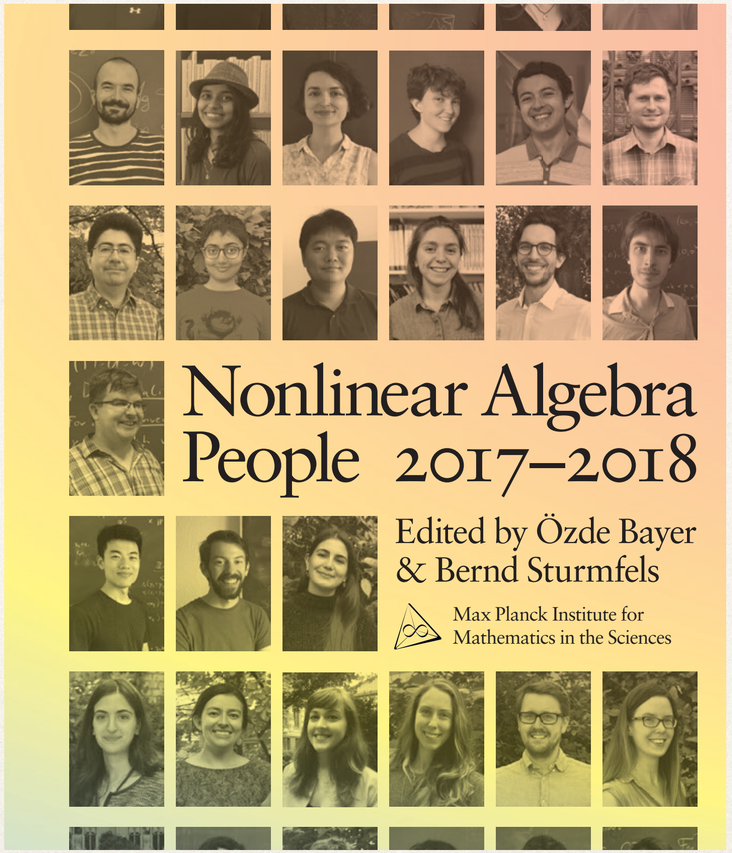
2011-2016

Postdoc: Brown University



2016-2017

Postdoc: Max Planck Institute



Tenure-track positions

- Wesleyan University
- Bentley University

Gender bias **exists** if women and men receive different evaluations that cannot be explained by objective differences in teaching quality.

Importance

Student evaluations are a frequently used assessment criterion for faculty performance in academia. They are often part of hiring, tenure, and promotion decisions and, thus, have a strong impact on career progression.

GENDER BIAS IN TEACHING EVALUATIONS

Friederike Mengel
University of Essex and Lund University

Jan Sauermann
Swedish Institute for Social Research (SOFI), Stockholm University

Ulf Zölitz
University of Zurich

Abstract
This paper provides new evidence on gender bias in teaching evaluations. We exploit a quasi-experimental dataset of 19,952 student evaluations of university faculty in a context where students are randomly allocated to female or male instructors. Despite the fact that neither students' grades nor self-study hours are affected by the instructor's gender, we find that women receive systematically lower teaching evaluations than their male colleagues. This bias is driven by male students' evaluations, is larger for mathematical courses, and particularly pronounced for junior women. The gender bias in teaching evaluations we document may have direct as well as indirect effects on the career progression of women by affecting junior women's confidence and through the reallocation of instructor resources away from research and toward teaching. (JEL: J16, J71, I23, J45)

Data

- Collected at the School of Business and Economics (SBE) of Maastricht University in the Netherlands and spans the academic years 2009/2010 to 2012/2013, including all bachelor and master programs.
- 735 different instructors, 9,010 students, 809 courses, and 6,206 sections.
- Students are randomly assigned to section instructors within courses. (this helps us to overcome selection problems)
- The data contains both a detailed set of students' subjective course evaluation items as well as their course grades. (to link objective performance indicators to subjective evaluation outcomes at the individual level)
- The data contains information on self-reported study hours. (measure of effort students put into the course)

TABLE 1. Descriptives statistics—full sample and estimation sample.

	(1) Full sample	(2) Estimation sample
Female instructor	0.348	0.344
	(0.476)	(0.475)
Female student	0.376	0.435
	(0.484)	(0.496)
Evaluation participation	0.363	1.000
	(0.481)	(0.000)
Course dropout	0.073	0.000
	(0.261)	(0.000)
Grade (first sit)	6.679	6.929
	(1.795)	(1.664)
GPA	6.806	7.132
	(1.202)	(1.072)
Dutch	0.302	0.278
	(0.459)	(0.448)
German	0.511	0.561
	(0.500)	(0.496)
Other nationality	0.148	0.161
	(0.355)	(0.367)
Economics	0.279	0.256
	(0.448)	(0.436)
Business	0.537	0.593
	(0.499)	(0.491)
Other study field	0.184	0.152
	(0.388)	(0.359)
Master student	0.247	0.303
	(0.431)	(0.460)
Age	20.861	21.077
	(2.268)	(2.305)
Overall number of courses per student	17.007	17.330
	(8.618)	(8.145)
Section size	13.639	13.606
	(2.127)	(2.061)
Section share female students	0.382	0.391
	(0.153)	(0.157)
Course-year share female students	0.380	0.386
	(0.089)	(0.093)
Observations	75,330	19,952
Number of students	9,010	4,848
Number of instructors	735	666

Evaluations

- In the last teaching week before the final exams, students receive an email with a link to the online teaching evaluation, followed by a reminder a few days later. Participation in the evaluation survey is only possible before the exam takes place.

- Likewise, faculty members receive no information about their evaluation before they have submitted the final course grades to the examination office.

- Evaluation survey: instructor-related statements (five items), group-related statements (two items), course material-related statements (five items), and course-related statements (four items). (Course materials are centrally provided by the course coordinator and are identical for all section instructors. All evaluation questions except study hours are answered on a five point Likert scale.)

TABLE 4. Evaluation items.

	(1) Mean	(2) Stand. Dev.
Instructor-related questions		
"The teacher sufficiently mastered the course content" (T1)	4.282	0.977
"The teacher stimulated the transfer of what I learned in this course to other contexts" (T2)	3.893	1.119
"The teacher encouraged all students to participate in the (section) group discussions" (T3)	3.551	1.209
"The teacher was enthusiastic in guiding our group" (T4)	4.022	1.125
"The teacher initiated evaluation of the group functioning" (T5)	3.595	1.247
Average of teacher-related questions	3.871	0.927
Group-related questions		
"Working in sections with my fellow-students helped me to better understand the subject matters of this course" (G1)	3.950	0.958
"My section group has functioned well" (G2)	3.943	0.962
Average of group-related questions	3.947	0.853
Material-related questions		
"The learning materials stimulated me to start and keep on studying" (M1)	3.425	1.131
"The learning materials stimulated discussion with my fellow students" (M2)	3.633	1.015
"The learning materials were related to real life situations" (M3)	3.933	0.971
"The textbook, the reader and/or electronic resources helped me studying the subject matters of this course" (M4)	3.667	1.067
"In this course EleUM has helped me in my learning" (M5)	3.110	1.073
Average of material-related questions	3.572	0.800
Course-related questions		
"The course objectives made me clear what and how I had to study" (C1)	3.467	1.074
"The lectures contributed to a better understanding of the subject matter of this course" (C2)	3.198	1.255
"The course fits well in the educational program" (C3)	4.020	0.995
"The time scheduled for this course was not sufficient to reach the block objectives" (C4)	3.151	1.234
Average of course-related questions	3.476	0.721
Study hours		
"How many hours per week on the average (excluding contact hours) did you spend on self-study (presentations, cases, assignments, studying literature, etc.)?"	14.07	8.071

Notes: Except for the number of study hours, all items are answered on a Likert scale from 1 ("very bad"), over 3 ("sufficient") to 5 ("very good"). Statistics are calculated for the estimation sample ($N = 19{,}952$). Missing values of sub-questions are not considered for the calculation of averages. EleUM stands for Electronic Learning Environment at Maastricht University.

Out of the full sample of all student-course registrations, 36% participate in the instructor evaluation.[16] This creates the potential for sample selection bias. Column (2)

16. If we require non-missing values for GPA among those who respond, we only observe 26% of the total sample (where the total sample includes those where GPA is missing).

TABLE 5. Gender bias in students' evaluations.

Dependent variable	(1) Instructor-related	(2) Group-related	(3) Material-related	(4) Course-related
Female instructor (β_1)	−0.2069***	−0.0579**	−0.0570**	−0.0780***
	(0.0310)	(0.0260)	(0.0231)	(0.0229)
Female student (β_2)	−0.1126***	−0.0121	−0.0287	−0.0373**
	(0.0184)	(0.0190)	(0.0178)	(0.0174)
Female instructor × Female student (β_3)	0.1309***	0.0493	0.0265	0.0635**
	(0.0326)	(0.0315)	(0.0297)	(0.0293)
Grade (first sit)	0.0253***	0.0221***	0.0442***	0.0528***
	(0.0058)	(0.0059)	(0.0058)	(0.0058)
GPA	−0.0633***	−0.0659***	−0.0377***	−0.0227***
	(0.0089)	(0.0088)	(0.0084)	(0.0083)
German	−0.0204	0.0129	0.0096	−0.0518***
	(0.0183)	(0.0186)	(0.0175)	(0.0177)
Other nationality	0.1588***	0.1162***	0.2418***	0.0871***
	(0.0220)	(0.0228)	(0.0222)	(0.0218)
Economics	−0.0989**	−0.0116	−0.0688	−0.1768***
	(0.0500)	(0.0534)	(0.0510)	(0.0529)
Other study field	−0.0777	−0.1264	−0.0566	0.0031
	(0.0840)	(0.0841)	(0.0806)	(0.0724)
Age	0.0138***	−0.0141***	0.0037	0.0064
	(0.0045)	(0.0047)	(0.0044)	(0.0045)
Section size	−0.0123	0.0009	−0.0047	−0.0106
	(0.0090)	(0.0080)	(0.0071)	(0.0071)
Constant	−0.1065	−0.0021	0.4323	−0.4096
	(0.4320)	(0.3165)	(0.3339)	(0.4434)
Observations	19,952	19,952	19,952	19,952
R-squared	0.1961	0.1559	0.2214	0.2360
$\beta_1 + \beta_3$	−0.0760**	−0.00855	−0.0305	−0.0145
	(0.0349)	(0.0292)	(0.0250)	(0.0244)

Notes: All regressions include course fixed effects and parallel course fixed effects for courses taken at the same time. Robust standard errors clustered at the section level in parentheses. All independent variables refer to student characteristics. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Results

- Male students evaluate female instructors 20.7% of a standard deviation worse than male instructors. (0.2 points on a five point Likert scale)
- Female students evaluate female instructors 7.6% of a standard deviation worse compared to male instructors.
- In a setting where 50% of students are female and 50% male, the male instructor would receive a 14.2% of a standard deviation higher evaluation than his female colleague.

Implications

Lower ratings for female instructors translate into substantial differences in rankings based on gender, which could manifest in other outcomes that are (partially) influenced by these rankings.

Concrete example: Teaching Awards. At the SBE in teaching awards are given in three categories (student instructors, undergraduate teaching, and graduate teaching). The share of female teaching instructors in the three categories is 40%, 38%, and 32%, respectively, and the share of female instructors among nominees is 15%, 26%, and 27%. There might be other reasons that cause this under-representation of women among nominees. However, these numbers are in line with the findings showing that female instructors receive substantially lower teaching evaluations compared to their male colleagues.

TABLE 7. Effect of instructor gender on instructor evaluation by seniority level.

	→ Increasing Seniority Instructors →				
	Student	Ph.D. student	Lecturer	Professor	Overall
Male students (β_1)	−0.2379***	−0.2798***	−0.0392	0.085	−0.2069***
	(0.0642)	(0.077)	(0.0619)	(0.1266)	(0.031)
Female students ($\beta_1 + \beta_3$)	−0.274***	−0.1359	0.1232*	0.2583**	−0.076**
	(0.0709)	(0.0862)	(0.0721)	(0.1179)	(0.0349)
Observations	5,352	4,801	5,700	4,099	19,952
R-squared	0.2839	0.3261	0.239	0.4473	0.1961

Notes: Dependent variable: Instructor evaluation. All estimates are based on regressions that include course fixed effects, parallel course fixed effects for the courses taken at the same time, section size and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level are in parentheses. The full table with student seniority can be found in Table B.12 in the Online Appendix. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Which Instructors are Subject to Gender Bias?

- Female student instructors receive 24% of a standard deviation worse ratings than their male colleagues if they are rated by male students. Remarkably, female students rate junior instructors very low as well. Junior female instructors receive evaluations that are 13.6%–27.4% of a standard deviation lower if they are rated by female students.

The result that predominantly junior women are subject to the bias implies that two otherwise comparable female and male job candidates would go on the market with a significantly different teaching portfolio.

TABLE 7. Effect of instructor gender on instructor evaluation by seniority level.

	→ Increasing Seniority Instructors →				
	Student	Ph.D. student	Lecturer	Professor	Overall
Male students (β_1)	−0.2379***	−0.2798***	−0.0392	0.085	−0.2069***
	(0.0642)	(0.077)	(0.0619)	(0.1266)	(0.031)
Female students ($\beta_1 + \beta_3$)	−0.274***	−0.1359	0.1232*	0.2583**	−0.076**
	(0.0709)	(0.0862)	(0.0721)	(0.1179)	(0.0349)
Observations	5,352	4,801	5,700	4,099	19,952
R-squared	0.2839	0.3261	0.239	0.4473	0.1961

Notes: Dependent variable: Instructor evaluation. All estimates are based on regressions that include course fixed effects, parallel course fixed effects for the courses taken at the same time, section size and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level are in parentheses. The full table with student seniority can be found in Table B.12 in the Online Appendix. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Which Instructors are Subject to Gender Bias?

Female students, however, rate female professors 25.8% of a standard deviation higher than male professors.

- One interpretation: seniority conveys a sense of authority to women that junior instructors lack.
- An alternative explanation: only the best female instructors "survive" the competition and reach the professor level. Thus, the only reason they receive similar ratings compared to their male counterparts is that they are actually much better teachers. (Data about student effort (study hours) and student grades according to the gender and seniority of the instructor does not support the idea that senior female instructors affect student outcomes positively.)

Math vs No-Math

- When female instructors teach courses with mathematical content, they risk being judged by the negative stereotype that women have weaker math ability.

- Male students rate female instructors around 32% of a standard deviation lower than they rate male instructors in these courses. For female students the effect is also large: female students rate female instructors in math-related courses around 28% of a standard deviation lower than they rate male instructors in these courses.

TABLE 10. Effect of instructor gender on instructor evaluation, study hours, and grades—by course content.

	(1)	(2)	(3)	(4)	(5)	(6)
	Instructor evaluation		Study hours		Grade	
Course content	No math	Math	No math	Math	No math	Math
Female instructor (β_1)	−0.1717***	−0.3197***	0.0192	0.1372	0.0170	0.0308
	(0.0329)	(0.0847)	(0.1925)	(0.3919)	(0.0357)	(0.0516)
Female student (β_2)	−0.1063***	−0.1488***	1.3544***	1.2709***	0.0174	−0.1225***
	(0.0216)	(0.0380)	(0.1767)	(0.2800)	(0.0276)	(0.0374)
Female instructor × Female student (β_3)	0.1366***	0.0421	−0.0700	−0.2207	0.0433	−0.1071
	(0.0356)	(0.0867)	(0.2754)	(0.5437)	(0.0468)	(0.0769)
Constant	1.0299***	0.1286	4.6886	8.6955*	−0.0429	0.9692
	(0.3507)	(0.5265)	(4.3592)	(4.5853)	(0.7119)	(0.7809)
Observations	14,843	4,820	14,843	4,820	14,843	4,820
R-squared	0.1851	0.2239	0.2682	0.2477	0.4730	0.6100
$\beta_1 + \beta_3$	−0.0351	−0.278***	−0.0508	−0.0835	0.0603*	−0.0763
	(0.0380)	(0.0903)	(0.229)	(0.406)	(0.0353)	(0.0590)

Notes: All regressions include course fixed effects, parallel course fixed effects for the courses taken at the same time, section size, and other control variables for students' characteristics (GPA, grade, nationality, field of study, age). Robust standard errors clustered at the section level are in parentheses. "Math" courses are defined as courses where courses require or explicitly contain math or statistics prerequisites, according to the course description. $^{*}p < 0.1$; $^{***}p < 0.01$.

Conclusion

- Female instructors receive systematically lower evaluations from both female and male students.
- Evaluating women worse is more pronounced among male students.
- Junior female instructors and those in math related courses consistently receive lower evaluation scores.
- No evidence that these differences are driven by gender differences in teaching skills. The results show that the gender of the instructor does not affect current or future grades nor does it impact the effort of students, measured as self-reported study hours.

Further Reading?

- A US study conducted an experiment whereby the instructors of an online course operated under two differently gendered avatars. This research found that students rated the male avatar significantly higher than the female avatar, regardless of the instructor's actual gender, but the study was based on a sample size of 43 students assigned to 4 different instructors.

SpringerLink

Published: 05 December 2014

What's in a Name: Exposing Gender Bias in Student Ratings of Teaching

Lillian MacNell, Adam Driscoll & Andrea N. Hunt

Innovative Higher Education **40**, 291–303 (2015) | Cite this article

28k Accesses | **347** Citations | **730** Altmetric | Metrics

Abstract

Student ratings of teaching play a significant role in career outcomes for higher education instructors. Although instructor gender has been shown to play an important role in influencing student ratings, the extent and nature of that role remains contested. While difficult to separate gender from teaching practices in person, it is possible to disguise an instructor's gender identity online. In our experiment, assistant instructors in an online class each operated under two different gender identities. Students rated the male identity significantly higher than the female identity, regardless of the instructor's actual gender, demonstrating gender bias. Given the vital role that student ratings play in academic career trajectories, this finding warrants considerable attention.

Recommendations for Better Evaluation

There are many well-documented ways that measurement bias and equity bias shape student evaluations of teaching. Given these biases and their role in personnel decisions, many colleges and universities – sometimes at the behest of faculty unions – are reevaluating how these ratings are used. We argue that we need not "throw out the baby with the bathwater" and eliminate the role of student evaluations entirely. Rather, they should be properly contextualized and used with caution. We offer the following six actions that individual faculty members and universities can take to make the use of student ratings more responsible.

1. Contextualize evaluations as perceptions of student learning, not as a measure of actual teaching.

   Although they are commonly called "student evaluations of teaching," SETs do not actually evaluate teaching. Instead, student evaluations represent their perception or experiences in a course (Linse, 2017; Abrami, 2001; Arreola, 2004). Students should not, and arguably cannot, evaluate teaching. A more accurate name for these experiences would be student experience questionnaires or student perceptions of learning. When properly contextualized as feedback on experience, rather than evaluating teaching, these assessments can provide useful feedback for faculty and administrators.
2. Be proactive about increasing the validity of the assessment by improving response rates.

   Administrators should not distribute or use assessments based on a low response rate. A low response rate makes it more likely that the sample is unrepresentative, which calls into question the validity of the assessment (Chapman & Joines, 2017; Adams & Umbach, 2012). Faculty can improve the response rate on evaluations by providing time for students to complete them in class, even if the evaluations are distributed online. Faculty should leave the room while this takes place, to alleviate pressure on students. Faculty can also improve the response rate by discussing the purpose of evaluations and how they can lead to improvements in the course for future students (Linse, 2017; Chapman & Joines, 2017).
3. Administrators should interpret the results of student ratings with caution.

   Student evaluations are not designed to be used as a comparative metric across faculty (Franklin, 2001); rather, their purpose is to gather information about how students perceived a faculty member teaching a certain course. As such, evaluations should be used to compare a faculty member's trajectory of teaching over time, and ideally, within a single course (Linse, 2017). Because one way that equity bias manifests is through lower evaluations for astereotypic instructors (i.e., women in male-dominated fields and vice versa), comparisons across faculty members further disadvantage already marginalized faculty.

Journal of Academic Ethics (2022) 20:73–84
https://doi.org/10.1007/s10805-021-09400-w

Check for updates

Evaluating Student Evaluations of Teaching: a Review of Measurement and Equity Bias in SETs and Recommendations for Ethical Reform

Rebecca J. Kreitzer[1] · Jennie Sweet-Cushman[2]

Accepted: 27 January 2021 / Published online: 9 February 2021

Abstract
Student evaluations of teaching are ubiquitous in the academe as a metric for assessing teaching and frequently used in critical personnel decisions. Yet, there is ample evidence documenting both measurement and equity bias in these assessments. Student Evaluations of Teaching (SETs) have low or no correlation with learning. Furthermore, scholars using different data and different methodologies routinely find that women faculty, faculty of color, and other marginalized groups are subject to a disadvantage in SETs. Extant

How to improve the system?