Student evaluations of teaching are not only unreliable, they are significantly biased against female instructors.

A series of studies across countries and disciplines in higher education confirm that student evaluations of teaching (SET) are significantly correlated with instructor gender, with students regularly rating female instructors lower than male peers. **Anne Boring**, **Kellie Ottoboni** *and* **Philip B. Stark** *argue the findings warrant serious attention in light of increasing pressure on universities to measure teaching effectiveness. Given the unreliability of the metric and the harmful impact these evaluations can have, universities should think carefully on the role of such evaluations in decision-making.*

Many universities rely heavily or exclusively on student evaluations of teaching (SET) for hiring, promoting and firing instructors. After all, who experiences teaching more directly than students? But to what extent do SET measure what universities expect them to measure—teaching effectiveness?

To answer this question, we apply nonparametric permutation tests to data from a natural experiment at a French university (the original study by Anne Boring is here), and a randomized, controlled, blind experiment in the US (the original study by Lillian MacNell, Adam Driscoll and Andrea N. Hunt is here). We confirm and extend the studies' main conclusion: Student evaluations of teaching (SET) are strongly associated with the gender of the instructor. Female instructors receive lower scores than male instructors. SET are also significantly correlated with students' grade expectations: students who expect to get higher grades give higher SET, on average. But SET are not strongly associated with learning outcomes.

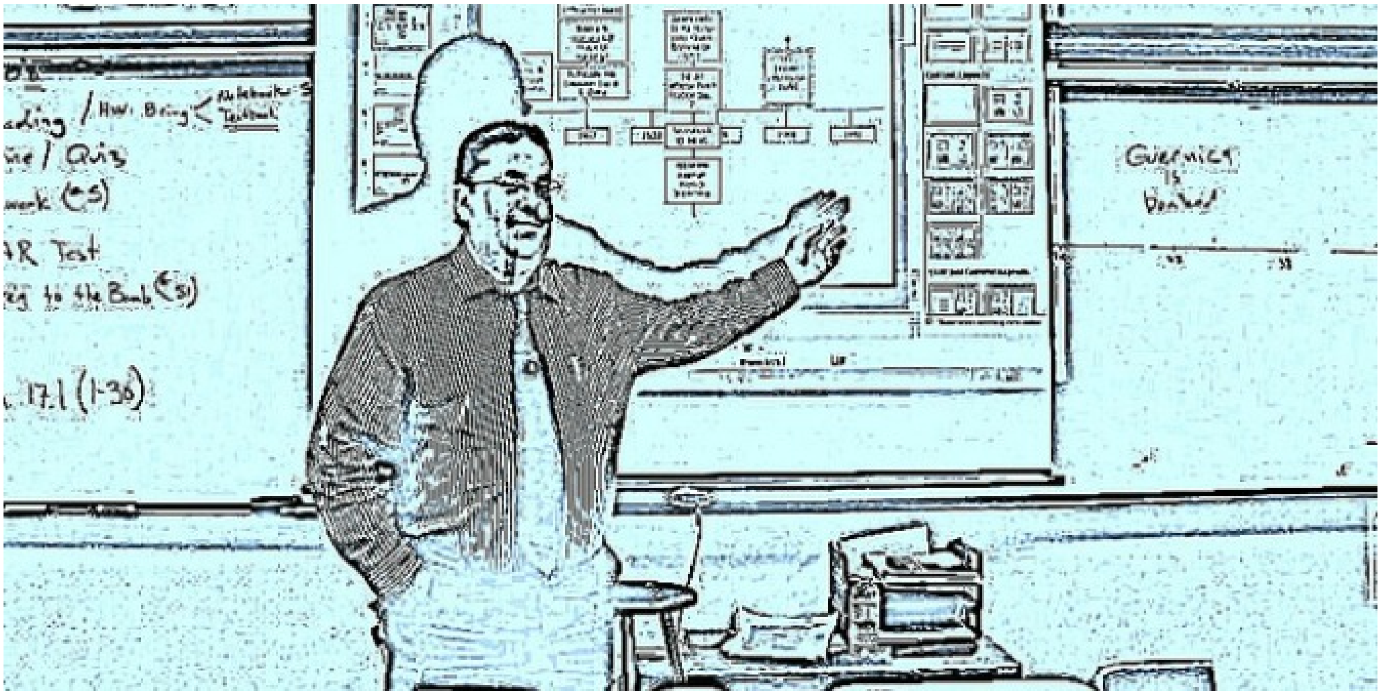
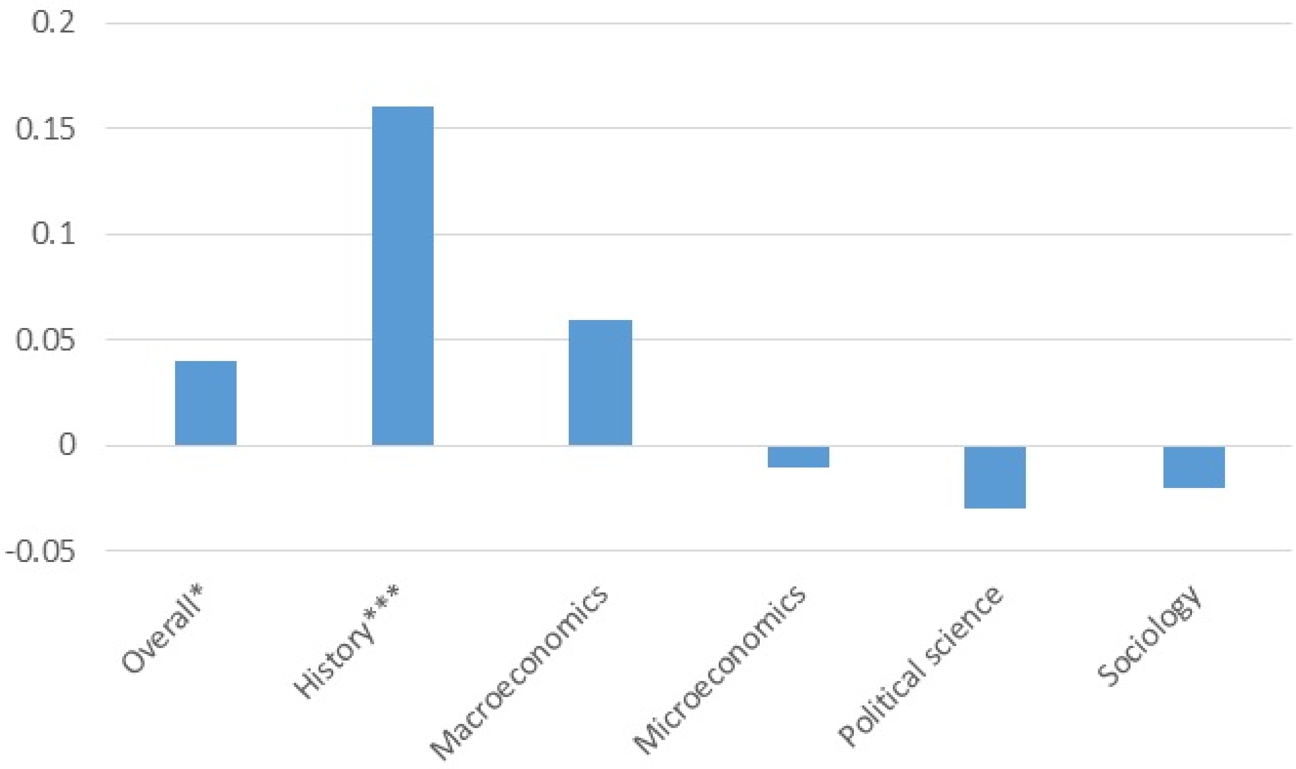


Image credit: Daniel R. Blume CC BY-SA (Flickr)

Some studies have found little difference between average SET for male and female instructors, but the design of those studies have serious flaws. Not only are they observational studies rather than experiments, they ask the wrong question, namely, “do male and female instructors get similar SET?” A better question is, “would female instructors get higher SET but for the mere fact that they are women?” We can answer that question using these unique data sets: “yes.”

The French Data: Since effective teaching should promote student learning, students of more effective instructors should have better learning outcomes on average. Students in different sections of each course, taught by different instructors, take the same final exam, allowing us to compare learning outcomes. We find that SET are at best weakly associated with student performance (Figure 1).

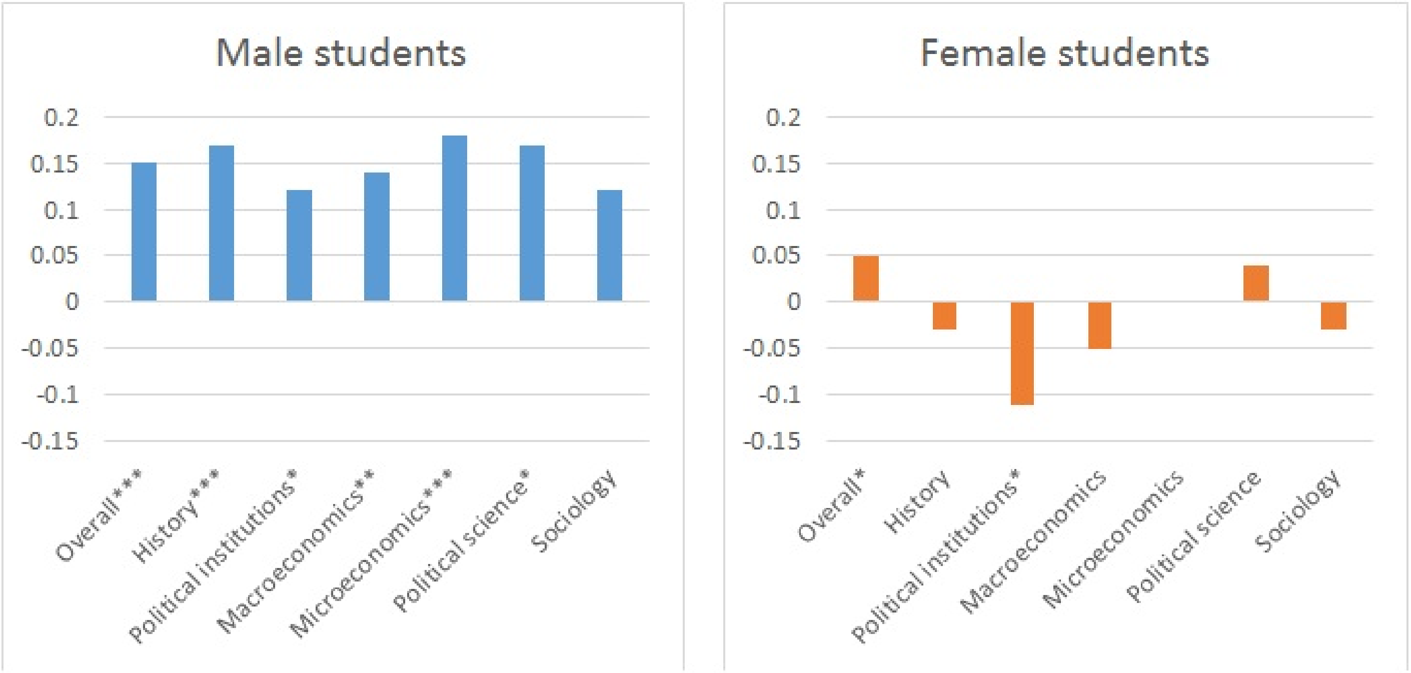
Figure 1. Average correlation between SET and final exam score, by subject



Note: p-values are one-sided, since, if SET measured teaching effectiveness, mean SET should be positively associated with mean final exam scores. Correlations are computed for course-level averages of SET and final exam score within years, then averaged across years. *** p<0.01, * p<0.1

On the other hand, SET are significantly correlated with instructor gender (male students gave higher SET to male instructors, Figure 2) and with students’ expected grades. This adds evidence to the hypothesis that instead of promoting better teaching, SET contribute to grade inflation. We find no evidence that male teachers are more effective than female teachers. If anything, students of male instructors perform worse on the final exam.

Figure 2. Average correlation between SET and gender concordance

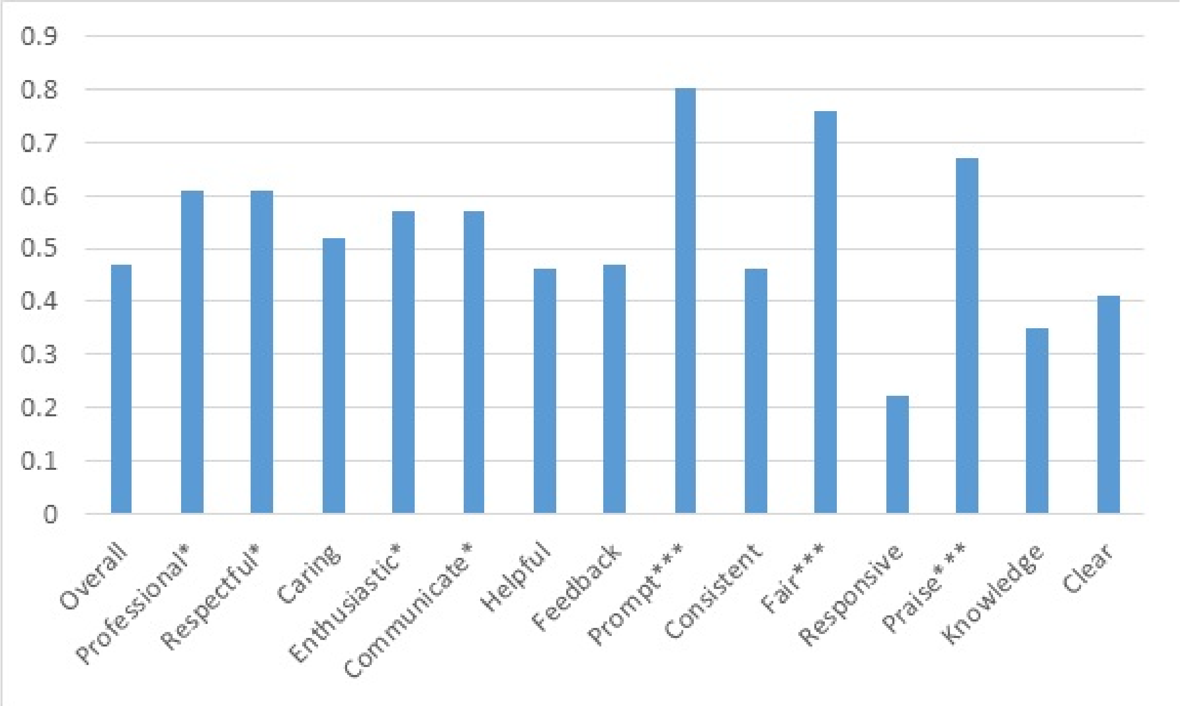


Note: p-values are two-sided. *** p<0.01, ** p<0.05, * p<0.1

The US Data: MacNell et al. (2014) collected data from four online sections of a course, two taught by a male instructor and two by a female instructor. Students were assigned randomly to the four sections. The male instructor taught one section using his own identity and switched identities with the female instructor for the other section, and vice versa. This lets us see how believing that an instructor is male or female affects SET for the very same instructor. We confirm the original authors' main finding that students generally rate perceived female instructors lower in several dimensions of teaching (Figure 3).

Even on measures one would expect to be objective, ratings were lower for perceived female instructors. For instance, graded assignments were returned simultaneously in all four sections, but students reported that the perceived female instructor was less prompt in returning assignments. Since SET were on a scale of 1 to 5, the observed difference in means, 0.80, is 20% of the full range.

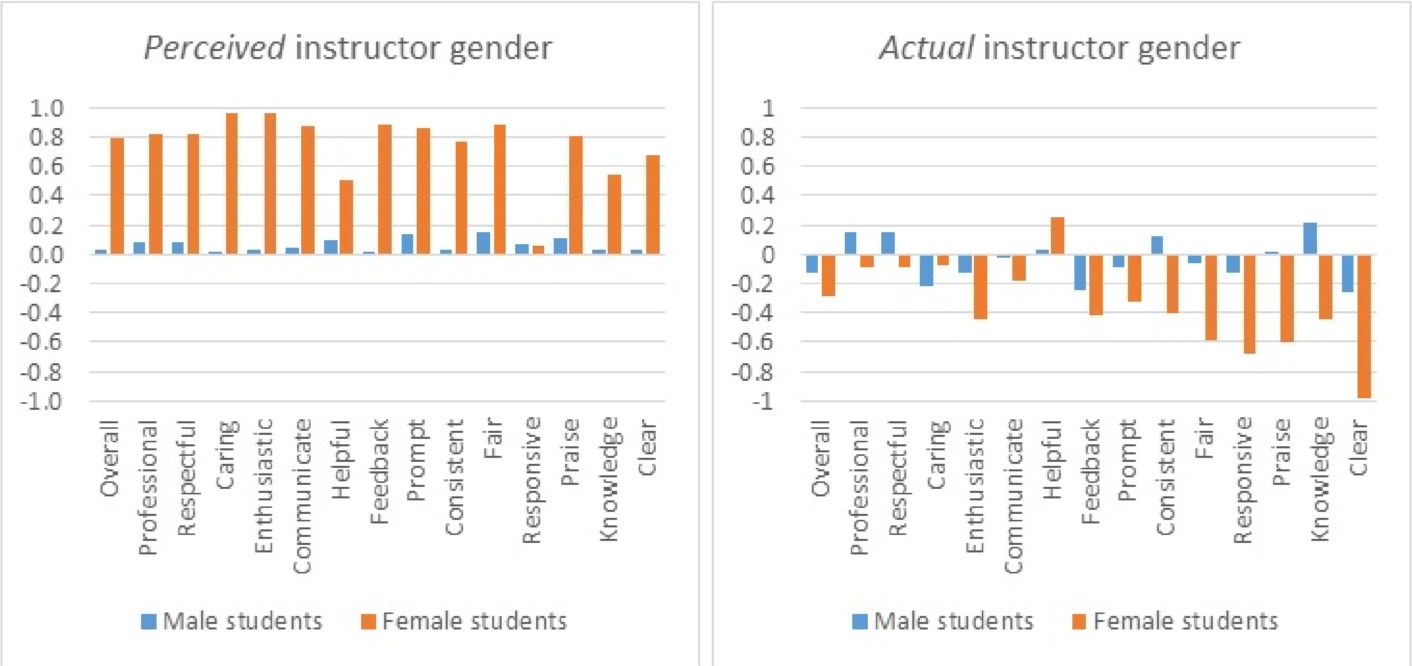
Figure 3. Difference in mean ratings and reported instructor gender (male minus female)



Note: The scale is 1-5 points, so a difference of 0.8 is 20% of the full range. p-values are two-sided. *** p<0.01, * p<0.1

In both the French and US data, male instructors got higher SET, but in the US data, female students tended to give higher scores to perceived male instructors (Figure 4), whereas in the French data, male students tended to give higher scores to male instructors.

Figure 4. Difference in mean SET by student gender, for perceived and actual instructor gender (male minus female)



Note: The p-values are not reported but can be found in thearticle (p.26-27).

In another study conducted in the Netherlands, researchers are finding that female instructors receive lower scores because male students give lower scores to female instructors. Differences among these studies might be cultural or related to topic, class size, mode of instruction (online versus face-to-face), ethnicity, race, physical attractiveness, or other confounding variables that have been found to affect SET. Clearly, there can be no simple adjustment for the bias.

The French data show that bias varies by course subject, further complicating any attempt to correct for these biases. The only field in which male students do not rate male instructors significantly higher is Sociology (Figure 1). This is especially interesting because Sociology is the only field in which there was near gender balance among instructors (46.4% female instructors). This might suggest that gender balance in a field affects gender stereotypes and might reduce bias against female instructors.

Why don't universities use better methods? SET are the familiar devil. Habits are hard to change. Alternatives (reviewing teaching materials, peer observation, surveying past students, and others) are more expensive and time-consuming, and this cost falls on faculty and administrators rather than on students. The mere fact that SET are numerical gives them an un-earned air of scientific precision and reliability. And reducing the complexity of teaching to a single (albeit meaningless) number makes it possible to compare teachers. This might seem useful to administrators, but it is a gross over-simplification of teaching quality.

The sign of any connection between SET and teaching effectiveness is murky, whereas the associations between SET and grade expectations and between SET and instructor gender are clear and significant. Because SET are evidently biased against women (and likely against other underrepresented and protected groups)—and worse, do not reliably measure teaching effectiveness—the onus should be on universities either to abandon SET for employment decisions or to prove that their reliance on SET does not have disparate impact.

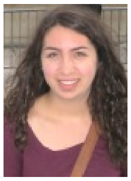
This blog post is based on a ScienceOpen preprint and can be found here: Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness DOI: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Featured image credit: The university and TAA negotiating an end to the 1970 strike (Wikimedia)

Note: This article gives the views of the author(s), and not the position of the LSE Impact blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

About the Authors:

Anne Boring is a research fellow at Sciences Po (OFCE-Presage) and a research affiliate at the University Paris Dauphine (LEDa-DIAL). Her main research interests include the study of gender biases and stereotypes in higher education. She also conducts research on interest groups, trade and development.



Kellie Ottoboni is a PhD student in the Statistics Department at the University of California, Berkeley and a fellow at the Berkeley Institute for Data Science. Her research interests include nonparametric statistics, causal inference, reproducibility, and applications in the health and social sciences.

Philip B. Stark is Professor of Statistics and Associate Dean of Mathematical and Physical Sciences at the University of California, Berkeley. He works primarily in uncertainty quantification, with applications to physical science, risk, natural disasters, elections, health, food security, litigation, and legislation.