Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness

Anne Boring, Kellie Ottoboni, Philip B. Stark

Draft April 17, 2015

The truth will set you free, but first it will piss you off.
Gloria Steinem

Abstract

We examine student evaluations of teaching (SET) at SciencesPo University, Paris, where all first-year students take the same courses (economics, history, political science, sociology, and political institutions). Students are assigned to sections of those courses as if at random, creating a natural experiment. Final exams are set for the entire course by the professor rather than the section instructor, and are graded anonymously. Hence, final exam scores are a proxy for the effectiveness of the section instructors. SET are mandatory. We study relationships among SET and the genders of students and instructors, topic, final exam scores, and students' grade expectations for 22,665 SETs of 372 instructors by 4,423 students over five years. Nonparametric permutation tests that aggregate within the 1,177 course sections show:

- the association between ratings and final exam scores is negative but insignificant (2-sided $P \approx 0.57$)
- the association between instructor gender and final exam scores is insignificant (students of male instructors do worse, 2-sided $P \approx 0.52$)
- the association between ratings and grade expectations is positive and highly significant (2-sided $P \approx 0.00$)
- the association between instructor gender and ratings is highly significant (men get higher ratings, 2-sided $P \approx 0.00$)
- male students rate male instructors significantly higher (2-sided $P \approx 0.00$) but male students score (insignificantly) lower on final exams in courses taught by male instructors (2-sided $P \approx 0.76$)
- female students rate male instructors higher, but not significantly (2-sided $P \approx 0.53$) but female students score (insignificantly) lower on final exams in courses taught by male instructors (2-sided ≈ 0.68)

These relationships vary by discipline.

1 Background

Student evaluations of teaching (SET) are used widely in higher education as a measure of teaching quality, and figure in the hiring, promotion, and firing of instructors, especially non-tenured faculty. SET are generally treated as a measure of teaching effectiveness, rather than, e.g., a measure of student satisfaction. Because ascertaining teaching effectiveness is so difficult—for students, faculty, and administrators alike—attempts to measure teaching effectiveness by surveying student opinion may suffer from conscious or unconscious biases. Recent work by **?** has demonstrated that this is the case: their randomized, controlled experiment shows that, on average, students rate a given instructor lower on every aspect of teaching (including "objective" measures such as timeliness) when they think the instructor is female than when they think the instructor is male.

Randomized, controlled experiments also show that SET do not measure teaching effectiveness; the key studies are **??**, which find that students confuse grades (or grade expectations) with long-term value.

Here, we use a remarkable census of SET by first-year students at SciencesPo (Paris) collected between 2008 and 2013, comprising 22,665 SETs by 4,423 students of 1,177 sections taught by 372 instructors. These data are discussed in detail by **?**. The key aspects of the data are these:

- All first year students take the same six courses, in history, macroeconomics, microeconomics, political institutions, political science, and sociology. Each course has one main professor, who delivers the lectures (to groups of approximately 900 students) and creates the final exams. Courses have many sections of 10–24 students. Those sections are taught by different instructors. The instructors have considerable pedagogical freedom.

- Students enroll in "triads" of sections of these courses. The enrollment process does not allow students the freedom to select individual instructors. The assignment of students to sections is "as if" at random.

- Section instructors provide interim grades during the term. Students know what their interim grades are, so interim grades are a good measure of grade expectations.

- Final exams are set by the main professor: all students in a given course take the same final. Final exams are graded anonymously in all

disciplines except Political Institutions (which we omit from analyses involving final exam scores). This makes performance on the final exam a reasonable measure of the value the section instructor adds: students of more effective instructors should do better on the final exam, on average.

- SET are mandatory: the response rates are nearly perfect.

SETs include closed-ended and open-ended questions, but the question that attracts the most attention is the overall score, which is considered to be a summary of the scores on the other questions.

We investigate hypotheses relating to whether the overall satisfaction score SET primarily measures teaching effectiveness or something else, for instance, the gender of the instructor or students' grade expectations. The data also allow us to determine whether there are systematic differences in how students rate courses in different disciplines.

We use nonparametric permutation tests rather than, for instance, logistic regression. Using nonparametric tests allows us to avoid counterfactual assumptions about generative models for the data, which regression-based methods (including ordinary linear regression, mixed effects models, logistic regression, etc.) and parametric methods such as t-tests and ANOVA, would require. The null hypotheses for our tests are simply that some characteristic—e.g., instructor gender—amounts to an arbitrary label, and might as well have been assigned at random.

Our analysis is conducted at the level of courses, which matches how SET are used in practice by institutions: typically, student responses in a given course are averaged, and those averages are compared across instances of the course, across courses in a department, and across departments within a university. Some of the statistical issues in this reduction of SET to averages are discussed by **?**

2 Tests

3 Code

Github repo. `https://github.com/kellieotto/SET-and-Gender-Bias`

4 Discussion

Implicit in the use of of SET as a Push back on the notion of "teaching effectiveness." There ought to be *some* interaction between characteristics of the instructor and those of the student. If "effectiveness" is intrinsic to the instructor, ratings in one class shouldn't depend on which other classes a student takes. Looking at ratings "per student" doesn't make sense if you are trying to measure some underlying platonic "effectiveness" intrinsic to the instructor. In particular, a showing that individual students who give a particular instructor higher ratings get higher grades, does not point to ....TO DO: FIX ME

Mention defenses of SET? The correlation between SET and performance isn't zero: it is positive, albeit not statistically significant. The larger point is that SET are better measures of student grade expectations and of instructor gender than they are of teaching effectiveness.

5 Conclusions