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

Anne Boring, Kellie Ottoboni, Philip B. Stark

Draft April 16, 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 significant (men get higher ratings, $P \approx 0.00$)
- 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. Student responses fail simple tests of data quality. For instance, 29% of students report spending impossible amounts of time on their courses.

1 Background

Student evaluations of teaching (SET) are widely used to assess 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, for instance, a measure of student satisfaction with the instructor. Because measuring teaching effectiveness objectively is so difficult—for students, faculty, and administrators alike—there is

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

Carrell & West, Braga, Paccagnella, & Pellizzari, 2011 on validity. Cite McNell, Boring on gender. Cite Lauer on comments. Cite defenders IDEA Benton & Cashin in defense of SET. Cite Spooren, Brockx, and Mortelmans 2013 for discussion on reliability and validity. Apart for student performance on the final exam, they cite an alternative to test for convergent validity (correlation between measures of student learning and SETs): using the class mean (as suggested by [**?**]).

How do these things fit together?

Reliability and validity: the correlation argument.

2 Data

2.1 SciencesPo Curriculum

First-year undergraduate students take six mandatory courses: in the fall, introduction to microeconomics, political institutions, and history; and in the spring, introduction to macroeconomics, political science, and sociology. Semesters last twelve weeks. Each week, for each course, students attend two one-hour lectures by a tenured professor (approximately 900 students per lecture) and two one-hour tutorial sections (between 10 and 24 students

per section). Our SET data include students' individual evaluations of instructors in these sections for microeconomics, history, political institutions, and macroeconomics for the five academic years 2008–2013, and for sociology and political science courses for the three academic years 2010–2013 (these two were introduced in 2010).

In the fall, students enroll in sections for the whole year, in cohorts called "triads": students in a triad take the same sections of all three courses. Students cannot pick and choose among individual instructors separately for each course, only among triads of sections. When possible according to students' and instructors' schedules, students stay in the same triad for the spring semester courses. Students are not allowed to change triads once courses have started in the fall.

The professor who delivers the lectures sets the content of the course and writes the final exam for the course as a whole. Section instructors devise their own syllabi; the administration encourages pedagogical freedom, especially regarding the assignments that instructors use to assess student progress during each term. Section instructors must, however, follow the course program set by the professor.

One way to measure teaching effectiveness is to check how students perform on the final exam. The main lecturer for the course writes the final exam, which all students have to take to pass the class (the final exam grade counts as one third of the final grade for the course). All papers are pooled and graded anonymously by different instructors of the course, such that instructors wouldn't know if they were actually grading one of their students or other students. There are written exams for five out of the six courses, whereas the political institutions final exam is an oral exam. The instructor who gives the final exam in the political institutions course is a different instructor from the one that students had during the semester; nonetheless, the grading is not anonymous and there is a large imbalance in the genders of the instructors in political institutions—52 men and 12 women–so we omit political institutions courses from our analysis. If SET scores actually measure teaching effectiveness, then students who rate instructors higher would be likely to obtain higher final grades.

Students have been completing their SETs online since 2008. The response rate is close to 100% as it is mandatory for students to complete their SETs. They risk a number of sanctions if they do not complete their SETs, such as not being able to register in the following semester. Students complete their SETs before they take their final exams. The SET scores are

anonymous to the teachers, who only have access to them once all grades have been officially recorded on student transcripts, several weeks after final exams. Instructors and academic coordinators then have access to SETs. When scores are low, the academic coordinator discusses the SETs with the instructor.

SETs include closed-ended and open-ended questions, but the question which attracts the most attention is the overall satisfaction score, which is considered to be a summary of the scores on the other questions. The complete set of questions that students fill-out is as follows:

How do you evaluate:

- the preparation and the organization of classes?
- the quality of the teaching materials?
- the clarity of the assessment criteria?
- the usefulness of feedback?
- your teacher's class leadership skills?
- your teacher's ability to encourage group work?
- your teacher's availability and communication skills?
- the course's ability to relate to current issues?
- your teacher's contribution to your intellectual development?

What is your overall level of satisfaction?

For this set of questions, students have a choice between answering non-pertinent, insufficient, average, good or excellent.

The other closed-ended questions that students answer are: Compared with other courses this semester, I invested much more effort / as much effort / much less effort in this course. How many assessments did you have throughout the semester? 0 to 2 / 3 to 4 / 5 to 6 / 7 or more Were written assignments given back within the time deadlines? Were oral presentation grades given back within the time deadlines?

Finally, the SET form ends with two closed-ended questions: What are the strong points of this course? What are the points that the teacher could improve?

We need to remove or run separate analyses for affirmative action (CEP) students. They tend to get lower overall grades than the rest of the students, which may be a confounder.

3 Tests

3.1 Per instructor

Pearson correlation between a summary statistic of effectiveness rating and a summary statistic of student performance, e.g., mean effectiveness (on various dimensions) and pass rate or mean final exam score. TO DO: CODE IS READY. DO WE DO ALL METRICS?

3.1.1 Gender

Pearson correlation between a summary statistic of effectiveness rating and gender of instructor TO DO: CODE IS READY. DO WE DO ALL METRICS?

3.2 Per student

For a single student, test correlation between course rating (overall, individual dimensions) and final grade/interim grade/professor gender. The null hypothesis is no correlation between rating and x.

The test assumes independence among students within a triad of classes. First null: all $3! = 6$ orderings equally likely. If we don't reject that, no need to go further.

We *could* allow for unequal probability that students have preferences related, e.g., to grades, but keep the independence assumption. One candidate weighting scheme is:

$$P(\text{teacher } i \text{ gets rated best}) = \frac{\%\text{ students given CAS } > t \text{ by teacher } i}{\sum_{j=1}^{3}(\%\text{ students given CAS } > t \text{ by teacher } j)}$$

Or more simply,

$$P(\text{teacher } i \text{ gets rated best}) = \frac{\text{CAS from teacher } i}{\sum_{j=1}^{3} \text{ CAS from teacher } j}$$

Aggregate the test statistics (Pearson correlation) across strata (different students) to get an overall p-value. There are roughly $(3!)^{14}$ different possible permutations. We expect to reject the null for the interim grades and for gender; do not expect to reject the null for final grades.

4 Inter-rater reliability

There is a distinction between teaching evaluations measuring something unique to each student (value added for them, from a particular teacher) versus measuring something intrinsic about the teacher. The goal of teaching evaluations is to measure intrinsic teaching ability. How well this is accomplished should be reflected in how similarly students rate their 6 teachers. We can rank the 6 teachers for each student based on the ratings they've assigned, then measure concordance between students in a triad by asking how often they ranked teacher i with rank j. Other measures of value added on the instructor level include the fraction of students who pass or the fraction of students with a final grade above x. On this line of reasoning, we can do a permutation test for the Pearson correlation between median rating from students in a class and the pass rate. Issues: We'll want to do these analyses separately for male and female students, since there seems to be an interaction effect between student gender and teacher gender. We assume stationarity: students will be the same and perform the same from semester to semester.

4.1 Gender effects

Look at the interaction of grades and gender: do students require higher grades from female teachers? Triplet effect for gender–pool triplets with same number of instructors of a given gender.

Another potential idea for controlling is to match. One approach is to use students as their own control, looking at a pair of classes in which they got the same grade, one taught by a male and the other by a female. An approach that will be more powerful and easier to implement is matching/binning students on their final exam scores/overall class grades; within matches or bins, we can do a sort of sign test to compare the mean ratings of male and female instructors. Within a bin, it's like a coin flip to decide whether the average male or female rating is higher - test if the coin is fair or biased. Preliminary

results suggest that the effect is more pronounced when comparing excellent to good ratings.

4.2 Relative or absolute

Hypothesis: students are comparing teachers rather than making absolute judgments.

5 Code

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

6 Discussion

7 Conclusions