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 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. Student responses fail simple tests of data quality. For instance, 29% of students report spending impossible amounts of time on their courses.