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

Kellie Ottoboni

Department of Statistics, UC Berkeley
Berkeley Institute for Data Science

May 3, 2016

Philip B. Stark
UC Berkeley

Anne Boring
SciencesPo University

Student evaluations of teachers (SET) are used to

- Quantify teaching effectiveness
- Compare instructors across courses
- Make hiring, firing, and promotion decisions

Are SET a valid measure of teaching effectiveness?

No!

We reanalyzed data from two studies:

- a natural experiment in France (Boring [2015])
- a randomized experiment in the US (MacNell et al. [2014])

Data from Boring [2015]

- Census of SET by first-year undergraduates, collected 2008–2013
- Students sign up for class times, don't know instructors; it's "as if" at random
- Male instructors are rated higher by male students than by female students
- SET correlate with grade expectations but not final grades

Data from MacNell et al. [2014]

- Students randomized to 4 online sections of a course
- In two sections, the TAs swapped identities
- Female-identified TA was rated lower on average in all categories

Characteristic	**F - M**
Overall	-0.47
Caring	-0.52
Consistent	-0.47
Enthusiastic	-0.57
Fair	-0.76
Feedback	-0.47
Helpful	-0.46
Knowledgeable	-0.35
Praise	-0.67
Professional	-0.61
Prompt	-0.80
Respectful	-0.61
Responsive	-0.22

Permutation tests

- Distribution-free tests, always give correct significance levels
- Compare assumptions:
  - Two-sample t-test: independent samples from normal distribution
  - Permutation test: randomization was fair
- Python package: permute
  `https://github.com/statlab/permute`

References

Anne Boring. Gender biases in student evaluations of teachers. Document de travail OFCE 13, OFCE, April 2015.

L. MacNell, A. Driscoll, and A. N. Hunt. Whats in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, pages 1–13, 2014.