What do Students Mean by "Teaching Effectiveness"?

Anne Boring, Kellie Ottoboni, Philip B. Stark

Draft April 7, 2015

1 Background

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.

How do these things fit together?

Reliability and validity: the correlation argument.

2 Data

TO DO: ANNE TO PROVIDE DESCRIPTION. TRIADS, ETC. 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.

Confidence bounds: We can lower bound the “female disadvantage,” i.e., how much a female teacher needs to improve her ratings in order to destroy the significant association from the test, assuming a constant effect size. Also can lower bound how much a teacher can lower the interim (continuous assessment) grades to break the association between interim grades and SET

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 for warm glow effect? 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.

4.3 Punishment

Look at students who took two courses from the same instructor, and got a lower course grade than the interim grade in the first course. Null hypothesis: equally likely to rate the instructor higher in the first and second course, independent of each other. Alternative: more likely to "punish" the instructor with a bad rating in the second course.

5 Code

Github repo.

6 Discussion

7 Conclusions