Evaluating Evaluations

Thanos Gentimis

November 9, 2016

1 Introduction

This is a brief document listing my ideas and findings about the use of STE (student teaching evaluations) as a measure of the quality of the professors and the instructional effectiveness. This effort started during the literature review for a paper we created to accompany the experimental teaching method of modules. During that time we were looking for objective measures to compare the method we were implementing versus the traditional methods used by other faculty. The idea of using the STE's was proposed and hastily discarded after only a few days due to the findings of the relevant literature.

It should be noted also that if we were to use my STE's as any measure of teaching effectiveness, the results will be fantastic, since I am currently on the top end of those, always way above the average in any category. And my opinion on that is ... that it doesn't matter at all. That measure has absolutely no connection to my performance as a professor, nor to the retention of the material by the students in my classroom. According to personal research, some important measures like student's average grade, student's retention in higher level classes are not at all correlated to my evaluations each year (with anecdotal evidence communicated through other Calculus I professors). The only one measure that I saw a difference is a negative correlation between students dropping in my class vs other classes. But again the factors here are unclear.

Still I would like to share with you our findings from the relevant bibliography search we did and which is available at the end of this manuscript. I propose also, that the relevant office conducts a statistical analysis connecting: average grades, drop rate, and retention in higher classes with the professors STE. I bet the results will be as clear as the bibliographic sources I am citing here, meaning no clear correlation present.

2 Literature Review

In his paper an Evaluation of Course Evaluations P. Stark clearly states: "*The common practice of relying on averages of student teaching evaluation scores as the primary measure of teaching, effectiveness for promotion and tenure decisions should be abandoned for substantive and statistical reasons: There is strong evidence that student responses to questions of effectiveness do not measure teaching effectiveness. Response rates and response variability matter. And comparing averages of categorical responses, even if the categories are represented by numbers, makes little sense. Student ratings of teaching are valuable when they ask the right questions, report response rates and score distributions, and are balanced by a variety of other sources and methods to evaluate teaching*" [5].

The first thing that student evaluations are plagued by is what in statistics is called "Response Bias". Ask yourselves, who would be prompted to complete a lengthy, optional survey about their professor and you already have your answer about the objectivity of your results. Couple that with the non-response and the whole process of using the sample of students that did respond to glean the truth about the ratings of the professor a text-book biased and unreliable sample.

As a matter of fact, I used this very example in my statistics class, to caution my students against surveys of that sort; It has all the textbook biases. Sometimes this bias works favorably for the professor, especially if he/she is more animated and makes a big deal out of the evaluations. A nice short speech about the importance of those evaluations especially before an exam can go a long way. Thus it is easy to bias your sample's opinion towards any direction.

Stark also mentions that: "*Personnel reviews routinely compare instructors' average scores to departmental averages. Such comparisons make no sense, as a matter of Statistics. They presume that the difference between 3 and 4 means the same thing as the difference between 6 and 7. They presume that the difference between 3 and 4 means the same thing to different students. They presume that 5 means the same thing to different students and to students in different courses.They presume that a 3 balance a 7 to make two 5's. For teaching evaluations, there is no reason any of those things should be true*"

In their paper "Student evaluations of teaching (mostly) do not measure teaching effectiveness", Anne Boring, Kellie Ottoboni, and Philip Stark show that:

1. SET are biased against female instructors by an amount that is large and statistically significant.

2. The bias affects how students rate even putatively objective aspects of teaching,

such as how promptly assignments are graded.

3. The bias varies by discipline and by student gender, among other things.

4. It is not possible to adjust for the bias, because it depends on so many factors.

5. SET are more sensitive to students' gender bias and grade expectations than they are to teaching effectiveness.

6. Gender biases can be large enough to cause more effective instructors to get lower SET than less effective instructors.

These findings are based on non-parametric statistical tests applied to two datasets: 23,001 SET of 379 instructors by 4,423 students in six mandatory first-year courses in a five year natural experiment at a French university, and 43 SET for four sections of an online course in a randomized, controlled, blind experiment at a US university [4].

The best way to reduce confounding is to assign students randomly to classes. That tends to mix students with different abilities and from easy and hard sections of the prequel across sections of sequels. This experiment has been done at the U.S. Air Force Academy (Carrell and West, 2010) [3] and Bocconi University in Milan, Italy (Braga, Paccagnella, and Pellizzari, 2014) [2]. These experiments found that teaching effectiveness, as measured by subsequent performance and career success, is negatively associated with SET scores. While these two student populations might not be representative of all students, the studies are the best we have seen. And their findings are concordant.

In their paper "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors" [3] Scott Carrell and James West note that: "*West Introductory course professors significantly affect student achievement in contemporaneous and follow-on related courses, but the effects are quite heterogeneous across subjects. Students of professors who as a group perform well in the initial mathematics course perform significantly worse in follow-on related math, science, and engineering courses. We find that the academic rank, teaching experience, and terminal degree status of mathematics and science professors are negatively correlated with contemporaneous student achievement, but positively related to follow-on course achievement. Across all subjects, student evaluations of professors are positive predictors of contemporaneous course achievement, but are poor predictors of follow-on course achievement.*"

A lot of research has also been done on the different biases that affect student evaluations. For example in [1] the authors propose that certain personality characteristics influence student evaluations of college professors. From the students points

of view, teacher-expressive characteristics such as warmth,enthusiasm, and extroversion apparently separate effective from ineffective teachers. They also report that there is evidence that college students rate male and female faculty according to subtle culturally conditioned age and gender stereotypes. The differences seem to occur more often on certain subjective, teacher-expressive factors and less often on objective content.

3 Conclusions

The previous statements make me believe that Student Teaching Evaluations are a measure of *something*. How does that *something* correlates to teaching effectiveness and retention of knowledge is unclear. Our colleagues, administration and students should be aware of the limitations of the STE to measure effectively and accurately. I will agree it it a good think to keep track but basing managerial decisions and creating records is just a cheap solution when it comes to evaluation.

I could offer alternatives (peer review, knowledge retention based on subsequent classes, Canvas usage and organization) which could create measures of the effectiveness of an educator. Yet, this short essay is already getting long so I prefer to stop here.

References

[1] Julianne Arbuckle and Benne D Williams. Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. *Sex Roles*, 49(9-10):507–516, 2003.

[2] Michela Braga, Marco Paccagnella, and Michele Pellizzari. Evaluating students evaluations of professors. *Economics of Education Review*, 41:71–88, 2014.

[3] Scott E. Carrell and James E. West. Does professor quality matter? evidence from random assignment of students to professors. *Journal Of Political Economy*, (118(3)):409–432, June 2010.

[4] Kellie Ottoboni, Anne Boring, and Philip Stark. Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, 2016.

[5] Philip B Stark and Richard Freishtat. An evaluation of course evaluations. *Center for Teaching and Learning, University of California, Berkley*, 2014.