

K. Melnick: SOME REMARKS ON STUDENT EVALUATIONS (2013)

Student end-of-term evaluations, together with peer review, are the main basis for evaluating faculty teaching for tenure and promotion at most US universities. While peer review may be skewed towards positive assessment of colleagues' teaching, student evaluations can be skewed upwards or downwards by a number of factors. Some of these factors are straightforward: small classes are rated more highly than big ones, morning classes more highly than afternoon, and elective, in-major classes are rated more highly than required classes in other departments. In the mathematics department, we teach many large-lecture courses that are required for other majors. Beyond these factors, there are further documented sources of bias that may be more insidious.

It is not clear that student evaluations correlate with learning. Three Ohio State economics faculty, B.A. Weinberg, B.M. Fleisher, and M. Hashimoto, attempted to study the relationship between learning, as measured by performance in a sequel course, and evaluation scores, after controlling for grades received in the course being evaluated. They found "student evaluations are positively related to current grades but unrelated to learning once current grades are controlled."¹ <http://www.nber.org/papers/w12844> Their data set comprises nearly 50,000 enrollments in almost 400 offerings of standard large economics courses at OSU. They also found no correlation between evaluations of a teacher and the portion of students of that teacher who continue to take more economics courses. In conclusion, they suggest that student evaluators may not be able to accurately assess how much they have learned in a course.

Some very fascinating studies indicate that evaluations of teachers may be based on non-verbal cues, and that end-of-semester evaluations are well predicted by surveys given after only 0.5 minutes of exposure to a silent video of a professor's lecture!^{2 3} An initial study by N. Ambady and R. Rosenthal involved nine undergraduate judges rating thirteen Harvard instructors, from a variety of disciplines, based on very short video clips. These results have since been reproduced on larger scales. It is interesting here that the authors presume student evaluations to be a good measure of teaching quality. They present their results as part of a body of work in social psychology on people's ability to make accurate judgments about others based on thin slices of expressive behavior.

Two Cornell psychology professors, W. M. Williams and S. J. Ceci, were inspired in part by this work to study these nonverbal effects while controlling for teaching quality.⁴ Ceci

¹B.A. Weinberg, B. M. Fleischer, and M. Hashimoto: "Evaluating methods for evaluating instruction: the case of higher education" *National Bureau of Economics Research Working paper 12844* (2007)

²N. Ambady and R. Rosenthal: Half a Minute: Predicting Teacher Evaluations from Thin Slices of Nonverbal Behavior and Physical Attractiveness *J. Personality and Soc. Psychol.* **64** 431-441 (1993)

³Nalini Ambady and Robert Rosenthal, Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis, *Psychol. Bull.* **111**, 256D274 (1992)

⁴W. M. Williams and S. J. Ceci: How'm I doing? *Change* **29** 12-23 (1997)

taught the same Human Development course, which he had been teaching for nearly 20 years, in subsequent semesters, with the same syllabus and lecture material; in the second semester he implemented purely stylistic changes recommended by a marketing consultant in a teaching workshop at Cornell. There were about 230 students in each semester, with similar demographics and similar grades in all aspects of the course. In the second semester, the instructor's mean rating increased by 0.84 on a scale from 0 to 5. Ratings in all categories increased significantly, including textbook quality, which rose by 0.92. Most importantly, the rating of amount learned increased by 1.12. These results, which have been reproduced in other studies, point to simple changes instructors can make to foster more positive feelings among students. The study also suggest that students may not be able to accurately quantify the amount they have learned in a course, without bias from their feelings about the teacher.

The Ambady-Rosenthal and Ceci-Williams studies together suggest that student evaluations are strongly influenced by a quickly formed, mostly subconscious impression of the teacher. In "Bias, the brain, and student evaluations of teaching," legal scholar D. J. Merritt asserts that "the behaviors that most influence these evaluations are rooted in physiology, culture, personality, and habit. Those behaviors are difficult for any faculty member to alter and they often reflect characteristics like race, gender, nationality, or socioeconomic class."⁵ There are many studies indicating that evaluations are negatively biased for instructors who are women or members of ethnic or racial minorities. The OSU economics study cited above, for example, included such findings, in spite of not detecting a correlation between instructor identity and the level of learning.

While there are certainly studies validating various aspects of student evaluations as a measure of teacher effectiveness, I try to maintain an awareness of how evaluation results may be influenced by psychological factors, factors that affect how all of us make judgments. I think we must also be conscious of how this measure of teaching quality advantages or disadvantages some of our colleagues. I look for additional ways to solicit student feedback in order to gain a more holistic picture of my teaching effectiveness, including midterm evaluations, conversations with my TAs, and conversations with students, both during and after the course.

⁵Deborah J. Merritt, Bias, the brain, and student evaluations of teaching, *St. John's Law Rev.* 82, 235-287 (2008)