Student Evaluations of Teaching Are Not Valid

BY JOHN W. LAWRENCE

SET scores are poor measures of teaching effectiveness and likely undermine educational standards.

In a review of the literature on student evaluations of teaching (SET), Philip B. Stark and Richard Freishtat—of the University of California, Berkeley, statistics department and the Center for Teaching and Learning, respectively—concluded, “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.” This is a startling conclusion, given that SET scores are the primary measure that many colleges and universities use to evaluate professors’ teaching. But a preponderance of evidence suggests that average SET scores are not valid measures of teaching effectiveness.
There are many statistical problems with SET scores. The response rate of student evaluations is often low. There is no reason to assume that the response pattern of those who do not complete the surveys would be similar to the pattern of those who do complete them. Some colleges assume that a low response rate is the professor's fault; however, no basis exists for this assumption. Also, average SET scores in small classes will be more greatly influenced by outliers, luck, and errors. SET scores are ordinal categorical variables in which participants make ratings that can be from poor (one) to great (seven). Stark and Freiheit point out that SET score numbers are labels, not values. We cannot assume the difference between one and two is the same as the difference between five and six. It does not make statistical sense to average categorical variables.

Even if SET score averages were statistically meaningful, it is impossible to compare them with other scores, such as the departmental average, without knowing the distribution of scores. For example, in baseball, if you don’t know the distribution of batting averages, you can’t know whether the difference between a .270 and .300 batting average is meaningful. Also, it makes no sense to compare SET scores of very different classes, such as a small physics course and a large lecture class on Shakespeare and hip-hop.

More problematic are the substantive concerns. SET scores are a poor measure of teaching effectiveness. They are correlated with many variables unrelated to teaching effectiveness, including the student’s grade expectation and enjoyment of the class; the instructor’s gender; race; age; and physical attractiveness; and the weather the day the survey is completed. In a 2016 study by economists Anne Boring and statisticians Kelli Ottoboni and Philip Stark, students in both France and the United States rated online professors more positively when they thought the instructor was a man. The researchers concluded that “SET are more sensitive to students’ gender bias and grade expectations than they are to teaching effectiveness.” In 2007, psychologists Robert Youmans and Benjamin Jee found that giving students chocolate before they completed teaching evaluations improved SET scores.

A 2016 study by two other psychologists, Nate Kornell and Hannah Hausman, reviewed the literature
If SET scores are such poor measures of teaching effectiveness and provide incentives for leniency, why do colleges and universities continue to use them? First, they are relatively easy to administer and inexpensive. Second, because they result in numerical scores, they have the appearance of being "objective." Third, the neoliberal zeitgeist emphasizes the importance of measuring individual performance instead of working as a community to address challenges such as improving teaching. And fourth, SET scores are part of a larger problem in higher education in which corporate administrators use assessment and credentialing procedures to exert control over faculty and students. The validity of the "assessments" is assumed and of secondary importance. In reality, the valid assessment of multifaceted phenomena such as teaching effectiveness is complex and requires scientifically rigorous investigations.

So, if SET scores are not measures of teacher effectiveness, how should colleges and universities evaluate teaching? Stark and Freishtat assert that measuring both teaching effectiveness and learning is complex and that we cannot do so reliably and routinely without applying rigorous experimental methods to all courses. They suggest that we focus on evaluating teaching. This requires evaluating the materials that professors create for classes and doing regular teaching observations. Yet, as a former department chair, I have read hundreds of peer teaching observations, and my impression is that the interrater reliability of peer observations is itself not particularly high.

Given the complexity of measuring good teaching, colleges and universities need to engage in this task with humility. It may be more fruitful for the institution to approach teaching observations as a strategy for facilitating collegial dialogue on striving for high-quality teaching and experimentation. The emphasis needs to be on mentoring, not potentially punitive evaluations. Moreover, in regard to improving student performance, institutional changes are likely more effective than focusing on individual professor performance. Particularly for minority and first-generation students, small class sizes and increased opportunities to interact with the professors and peers can improve a variety of outcomes.