An Evaluation of Course Evaluations

Philip B. Stark*

Department of Statistics, University of California, Berkeley
Berkeley, CA 94720, United States

Richard Freishtat

Center for Teaching and Learning, University of California, Berkeley
Berkeley, CA 94720, United States

*Corresponding author. E-mail: stark@stat.berkeley.edu

Draft: 2 June 2014

An evaluation of course evaluations

Student ratings of teaching have been used, studied, and debated for almost a century. This article examines student ratings of teaching from a statistical perspective. 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.

Keywords: student course evaluations, teaching evaluations, teaching effectiveness, statistics

Among faculty, course evaluations are a source of pride and satisfaction—and frustration and anxiety. High-stakes decisions including tenure and promotions rely on these evaluations. Yet it is widely believed that they are primarily a popularity contest; that it's easy to "game" the ratings; that good teachers get bad ratings and *vice versa*; and that fear of bad ratings stifles pedagogical innovation and encourages faculty to water down course content. What's the truth?

Background

Quantitative student ratings of teaching are the most common method to evaluate teaching (see, e.g., Cashin [1999]; Clayson [2009]; Davis [2009]; Seldin [1999]; Seldin, Miller, and Seldin [2010]). *De facto*, they define "effective teaching" for many purposes. They are popular partly because the measurement is easy: Students fill out forms. It takes about 10 minutes of class time and even less faculty time. The biggest cost is to transcribe the data; online evaluations automate that step. Averages of numerical student ratings have an air of objectivity simply because they are numerical. And comparing the average rating of any instructor to the average for her department as a whole is simple.

Since 1975, course evaluations at University of California, Berkeley have asked: Considering both the limitations and possibilities of the subject matter and course, how would you rate the overall teaching effectiveness of this instructor? Students respond on a Likert scale, ranging from 1 (not at all effective), to 4 (moderately effective), to 7 (extremely effective). In doing so, Berkeley invites those comparisons of average ratings. For instance, a sample letter by the U.C. Berkeley College of Letters and Sciences for department chairs to request a "targeted decoupling" of faculty salary says:

> Smith has a strong record of classroom teaching and mentorship. Recent student evaluations are good, and Smith's average scores for teaching effectiveness and course worth are (around) ____________ on a seven-point scale, which compares well with the relevant departmental averages.

We will argue that such comparisons are inappropriate and misleading.

Experimental evidence suggests that omnibus questions—such as "overall teaching effectiveness" and "course worth"—should be avoided entirely. The responses to such questions are strongly influenced by factors unrelated to learning, such as the gender, ethnicity, and attractiveness of the instructor. While students are in a good position to evaluate *some* aspects of teaching, there is compelling evidence that student evaluations are only tenuously connected to overall teaching effectiveness. They offer only a single perspective on a very complex and multifaceted teaching and learning process that no single source of evidence can reasonably evaluate. Defining and measuring teaching effectiveness are knotty problems, as we discuss below.

Here we examine student course evaluations from a statistical perspective. We argue that averages of rating scores should not even be calculated, much less compared across instructors, courses, or departments. Instead, frequency tables should be used to summarize scores. It is crucial to report survey response rates, not merely the number of respondents. Finally, we recommend complementary sources of evidence that can be combined with student teaching evaluations to provide more meaningful and reliable formative and summative assessments of teaching.

Statistics and Student Evaluations of Teaching

We start with a brief, nontechnical look at some statistical issues in collecting, summarizing, and comparing course evaluations. We illustrate some issues relying on "overall teaching effectiveness" scores, but we argue that those scores should not even be collected.

Who responds?

Some students are absent when in-class evaluations are administered. Some students who are present do not fill out the survey; similarly, some students do not fill out online evaluations of teaching. There are several ways to provide social and administrative incentives to students to encourage them to fill out online evaluations of teaching, for instance, allowing them to view their grades sooner if they have filled out the evaluations. The incentives, some of which have been tried elsewhere, have pros and cons. Conversely, some faculty express concerns that students who have not attended a single lecture can still fill out online evaluations. Regardless of the administration method and incentives, the *response rate* will likely be less than 100%. The lower the response rate, the less representative the responses might be. There is no justification for assuming that nonresponders are just like responders. Indeed, there is reason to think they are *not*: They were not present or they chose not to fill out the evaluation. Moreover, people tend to be motivated to act (e.g., fill out an online evaluation) more by anger than by satisfaction. Have you ever seen a public demonstration where people screamed, "we're content!"(see, e.g., http://xkcd.com/470/)? And differences between responders and nonresponders can matter.

As an extreme illustration, suppose that only half the class responds, and that all those "responders" rate the overall teaching effectiveness as 2/7. The average for the entire class might be as low as 1.5/7, if all the "nonresponders" would also have rated it 1/7. Or it might be as high as 4.5/7, if the nonresponders would have rated the effectiveness 7/7.

Berkeley's *Policy for the Evaluation of Teaching* (UC Berkeley 1987) requires faculty to provide an explanation if the response rate is below ⅔. This seems to presume that it is the instructor's fault if the response rate is low, and that a low response rate is in itself a sign of bad teaching. Is this fair or accurate? Consider three scenarios:
(1) The instructor has invested an enormous amount of effort in providing the material in several forms, including online materials, online self-test exercises, and webcast lectures; the course is at 8am. We might expect attendance and response rates to in-class evaluations to be low.
(2) The instructor is not following any text and has not provided notes or supplementary materials. Attending lecture is the only way to know what is covered. We might expect attendance and response rates to in-class evaluations to be high.
(3) The instructor is exceptionally entertaining, gives "hints" in lecture about what to expect on exams; the course is at 11am. We might expect attendance and response rates to in-class evaluations to be high.
The point: Response rates in themselves say little about teaching effectiveness. In reality, if the response rate is low, the data should not be considered representative of the class as a whole. An explanation solves nothing.

Averages of small samples are more susceptible to "the luck of the draw" than averages of larger samples. This can make teaching evaluations in small classes more extreme than evaluations in larger classes, even if the response rate is 100%. And students in small classes might imagine their anonymity to be more tenuous, perhaps reducing their willingness to respond truthfully or to respond at all.

Averages

Academic personnel review processes often invite comparing an instructor's average scores to the departmental average. Such averages and comparisons make no sense, as a matter of statistics. They presume that the difference between 3/7 and 4/7 means the same thing as the difference between 6/7 and 7/7. They presume that the difference between 3/7 and 4/7 means the same thing to different students. They presume that 5/7 means the same thing to different students and to students in different courses. They presume that a 3/7 "balances" a 7/7 to make two 5/7s. For teaching evaluations, there is no reason any of those things should be true (see, e.g., McCullough & Radson [2011]).

Effectiveness ratings are what statisticians call an *ordinal categorical* variable: The ratings fall in categories that have a natural order, from worst (1) to best (7). But the numbers are labels, not quantities. We could replace the numbers with descriptions and no information would be lost: The ratings might as well be "not at all effective," "slightly effective," and "extremely effective."

Does it make sense to take the average of "slightly effective" and "very effective"? Relying on average evaluation scores equates the effectiveness of an instructor who receives two ratings of 5/7 and the effectiveness of an instructor who receives a 3/7 and a 7/7, since both instructors have an average rating of 5/7. Are they really equivalent?

They are not, as this joke shows: Three statisticians go hunting. They spot a deer. The first statistician shoots; the shot passes a yard to the left of the deer. The second shoots; the shot passes a yard to the right of the deer. The third one yells, "We got it!"

Scatter matters

Comparing an individual instructor's average rating with the average rating for a course or a department is less informative than campus guidelines appear to assume. For instance, suppose that the departmental average for a particular course is 4.5/7, and the average for a particular instructor in a particular semester is 4.2/7. The instructor's rating is below average. How bad is that? Is the difference meaningful?

There is no way to tell from the averages alone, because of instructor-to-instructor and semester-to-semester variability. If all other instructors get an average of exactly 4.5/7 when they teach the course, 4.2/7 might be atypically low. On the other hand, if other instructors get 6/7 half the time and 3/7 the other half of the time, 4.2 is well within the spread of scores. Even if it made sense to average scores, the mere fact that one instructor's average rating is above or below the average for the department says very little. Instead of reporting averages, we should report the distribution of scores for instructors and for courses: the percentage of ratings that fall in each category (1–7). The distribution is easy to convey using a bar chart.

All the children are above average

At least half the faculty in any department will have average teaching evaluation scores at or below median for that department. Of course, it is possible for an entire department to be "above average" compared to an institutions' faculty as a whole. Rumor has it that department chairs sometimes argue in merit cases that a faculty member with below-average teaching evaluations is an excellent teacher—just perhaps not as good as the other teachers in the department, all of whom are superlative. This could be true in some departments, but it cannot be true in every department. With apologies to Garrison Keillor, all faculty at an institution cannot be above average for that institution.

Comparing incommensurables

Students' motivations for taking courses vary, in some cases systematically by the type of course or even with the particular course (e.g., prerequisites versus major electives). The nature of the interaction between students and faculty varies across types and sizes of course. These variations are large and may be confounded with teaching evaluation scores (see, e.g., Cranton & Smith, [1986], Feldman [1984, 1978]). Lower-division students and new transfer students have less experience with our courses here at Berkeley than seniors have. It is not clear how to make fair comparisons of student teaching evaluations across seminars, studios, labs, large lower-division courses, gateway courses, required upper-division courses, etc., although such comparisons are common (see, e.g., McKeachie [1997]).

Student Comments

Students are ideally situated to comment about *their experience* in the course, including factors that influence teaching effectiveness, such as the instructor's audibility, legibility, and availability outside class. They might also be able to judge clarity, but clarity may be confounded with the difficulty of the material. However, the depth and quality of students' comments vary widely by discipline. Students in STEM disciplines tend to write much less, and much less enthusiastically, than students in arts and humanities. That makes it hard to make fair comparisons across disciplines. Below are comments on two courses, one in Physical Sciences and one in Humanities. By the standards of the disciplines, all four student comments are glowing.

Physical Sciences:

> "Lectures are well organized and clear"
> "Very clear, organized and easy to work with"

Humanities:

> "There is great evaluation of understanding in this course and allows for critical analysis of the works and comparisons. The professor prepares the students well in an upbeat manner and engages the course content on a personal level, thereby captivating the class as if attending the theater. I've never had such pleasure taking a class. It has been truly incredible!"
> "Before this course I had only read 2 plays because they were required in High School. My only expectation was to become more familiar with the works. I did not expect to enjoy the selected texts as much as I did, once they were explained and analyzed in class. It was fascinating to see texts that the author's were influenced by; I had no idea that such a web of influence in Literature existed. I wish I could be more 'helpful' in this evaluation, but I cannot. I would not change a single thing about this course. I looked forward to coming to class everyday. I looked forward to doing the reading for this class. I only wish that it was a year long course so that I could be around the material, GSI's and professor for another semester."

While some student comments are extremely informative—and we strongly advocate that faculty read all student comments—it is not easy to compare comments fairly across disciplines (see, e.g., Cashin [1990]; Cashin & Clegg [1987]; Cranton & Smith [1986]; Feldman [1978]).

What Evaluations Measure

> If you can't prove what you want to prove, demonstrate something else and pretend that they are the same thing. In the daze that follows the collision of statistics with the human mind, hardly anybody will notice the difference.
> -D. Huff (1954)

To a great extent, this is what we do with student evaluations of teaching effectiveness. We do not measure teaching effectiveness. We measure what students say, and pretend it's the same thing. We dress up the responses by taking averages to one or two decimal places, and call it a day.

What is effective teaching? Presumably, it has something to do with learning. One definition is that an effective teacher is skillful at creating conditions conducive to learning. What is to be learned varies by discipline and by course: It might include facts, skills, ways of thinking, habits of mind, or a maturing of perspective. Some learning will happen no matter what the instructor does. Some students will not learn much no matter what the instructor does. How can we tell how much the instructor helped or hindered learning in a particular class?

What can we measure?

Measuring learning is hard: Course grades and exam scores are poor proxies, because courses and exams can be easy or hard. Beleche, Fairris & Marks (2012) point out:

> It is not clear that higher course grades necessarily reflect more learning. The positive association between grades and course evaluations may also reflect initial student ability and preferences, instructor grading leniency, or even a favorable meeting time, all of which may translate into higher grades and greater student satisfaction with the

course, but not necessarily to greater learning.

If someone other than the instructor set exams—the practice in some universities—we might be able to use exam scores to measure learning (see, e.g., http://xkcd.com/135/). But that is not how most universities work, and teaching to the test could be confounded with deep learning. Performance in follow-on courses and career success may be better measures, but time must pass to make such measurements, and it is difficult to track students over time. And how much of someone's career success can be attributed to a given course, years later?

A large amount of research on student evaluations of teaching that mostly addresses *reliability*: Do different students give the same instructor similar marks (see, e.g., Abrami et al. [2001]; Braskamp and Ory [1994]; Centra [2003]; Ory [2001]; Wachtel [1998]; Marsh and Roche [1997])? Would the same student give the same instructor the same mark later (see, e.g., Braskamp and Ory [1994]; Centra [1993]; Marsh [2007]; Marsh and Dunkin [1992]; Overall and Marsh [1980])? Reliability has little to do with whether evaluations measure effectiveness. A hundred bathroom scales might all report your weight to be the same. That does not mean the readings are accurate measures of your *height*—or even your weight, for that matter. Moreover, inter-rater reliability seems to be an odd thing to worry about, in part because it is easy to report the full distribution of student ratings, as advocated earlier. Scatter matters, and it can be measured *in situ* in every course.

Observational Studies versus Randomized Experiments

Most of the research on student teaching evaluations is based on *observational studies*. Students take whatever courses they choose from whomever they choose. The researchers watch and report. In the entire history of Science, there are few observational studies that justify inferences about causes—a notable exception being John Snow's research on the cause of cholera; his study amounts to a "natural experiment" (see http://www.stat.berkeley.edu/~stark/SticiGui/Text/experiments.htm#cholera for a discussion).

In general, to infer causes, such as whether good teaching results in good evaluation scores, requires a *controlled, randomized experiment*. In a controlled, randomized experiment, individuals are assigned to groups at random; the groups get different *treatments*; the outcomes are compared across groups to test whether the treatments have different effects and to estimate the sizes of those differences.

"Random" is not the same as "haphazard." In a randomized experiment, the experimenter deliberately uses a blind, non-discretionary chance mechanism to assign treatments to individuals. Randomization tends to mix individuals across groups in a balanced way. Differences in outcomes across groups are a combination of chance and differences in the treatments—and the chance mechanism is known. Absent randomization, differences among the groups other than the treatment can be *confounded* with the effect of the treatment, and there is no way to tell what portion of the observed differences is due to confounding (see, e.g., http://xkcd.com/552/). For instance, suppose that some students deliberately choose classes by finding the professor reputed to be the most lenient grader. Such students might then rate that professor highly for an "easy A." If those students choose sequel courses the same way, they are likely to get easy good grades in those as well, "proving" that the high ratings the first professor received were justified.

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 2008) and Bocconi University in Milan, Italy (Braga, Paccagnella, and Pellizzari 2011). These experiments confirm that good teachers can get bad evaluations: Teaching effectiveness, as measured by subsequent performance and career success, is negatively associated with teaching evaluations. While these two student populations might not be representative of university students broadly, the studies are among the best we have seen. And their findings are concordant.

What do student evaluations of teaching measure?

As mentioned before, student teaching evaluations are *reliable* in the sense that students often agree (Braskamp and Ory 1994; Centra 1993; Marsh 2007; Marsh and Dunkin 1992; Overall and Marsh 1980). But homogeneity of responses is red herring. It would be a rare instructor who was equally effective with students with different background, preparation, skill, disposition, maturity, and "learning style." That in itself suggests that if ratings are as consistent as some studies assert, perhaps ratings measure something other than teaching effectiveness: If a laboratory instrument always gives the same reading when its inputs vary substantially, it's broken.

If evaluations do not measure teaching effectiveness, what do they measure? While we do not vouch for any of the following studies, they reflect conflicting evidence and little consensus:

- Student teaching evaluation scores are highly correlated with students' grade expectations (Marsh and Cooper 1980; Short et al. 2012; Worthington 2002).
- Effectiveness scores and enjoyment scores are related. In a pilot of online course evaluations in the UC Berkeley Department of Statistics in fall 2012, among the 1486 students who rated the instructor's overall effectiveness and their enjoyment of the course on a 7-point scale, the correlation between instructor effectiveness and course enjoyment was 0.75, and the correlation between course effectiveness and course enjoyment was 0.8.
- Students' ratings of instructors can be predicted from the students' reaction to 30 seconds of silent video of the instructor: first impressions may dictate end-of-course evaluation scores, and physical attractiveness matters (Ambady and Rosenthal 1993).
- Gender, ethnicity, and the instructor's age matter (Anderson and Miller 1997; Basow 1995; Cramer and Alexitch 2000; Marsh and Dunkin 1992; Wachtel 1998; Weinberg et al. 2007; Worthington 2002).

Worthington (2002, 13) makes the troubling claim, "the questions in student evaluations of teaching concerning curriculum design, subject aims and objectives, and overall teaching performance appear most influenced by variables that are unrelated to effective teaching." U.C. Berkeley, like many universities, hangs its hat on just such a question about overall teaching performance.

What are student evaluations of teaching good for?

Students are in a good position to observe several aspects of teaching that *contribute* to effectiveness, such as clarity, pace, legibility, audibility, and availability. Student surveys can help get a picture of these things; the statistical issues raised in this article still matter (especially response rates, inappropriate use of averages, false numerical precision, and scatter). But, we should not ask students to rate teaching effectiveness *per se*. Students then tend to answer a rather different question from the question asked, regardless of their intentions. Calling the result a measure of teaching effectiveness does not make it one, any more than you can make a bathroom scale measure height by relabeling its dial "height." Calculating precise averages of such "height" measurements made with 100 different scales would not help.

A Better Way to Evaluate (and Improve) Teaching

Let's drop the pretense. We will never be able to measure teaching effectiveness reliably and routinely. It does not even seem possible to *define* teaching effectiveness quantitatively in a way that makes sense across disciplines. In some disciplines, measurement is possible but would require structural changes, for instance, assigning students to class sections at random and tracking their performance for semesters or years, or having different faculty set exams. And even that would not suffice in many disciplines.

If we want to understand how someone is teaching, we have to look. If we want to know what is going on the classroom, we have to look. If we want to know whether an instructor's materials are good, we have to look. Most of all, if we want teaching to improve, we have to pay attention to the teaching itself, not to the average of a list of student-reported numbers that bear a troubled and murky relationship to the teaching. We would benefit from visiting each other's classrooms and looking at others' teaching materials routinely. We can learn from each other, exchanging pedagogical ideas and practices.

When it's time to evaluate teaching, there are many things to rely on. We can look at student comments. We can look at materials the candidate created designing, redesigning, and teaching courses, such as syllabi, lecture notes, websites, textbooks, software, videos, assignments, and exams. We can look at faculty teaching statements. We can look at samples of student work, from homework to honors theses and dissertations. We can watch lectures. We can survey former students, advisees, and graduate teaching assistants. We can look at the job placement success of former graduate students. Etc.

We can look at all those things and ask: Is this a good and dedicated teacher? Is she engaged in teaching? Is she following pedagogical practices found to work in the discipline? Is she available to students? Is she putting in appropriate effort? Is she creating new materials, new courses, or new pedagogical approaches? Is she revising, refreshing, and reworking existing courses? Is she helping keep the curriculum in the department up to date? Is she trying to improve? Is she improving? Is she contributing to the university's teaching mission in a serious way? Is she supervising undergraduates for research, internships, and honors theses? Is she advising and mentoring graduate students? Is she serving on qualifying exam committees and thesis and dissertation committees? Do her students do well when they graduate?

Or is she checked out? Does she teach from the same stale lecture notes given to her two decades ago by a senior faculty member the first time she taught the course? Does she mumble in a monotone, facing the board, scribbling illegibly? Do her actions and demeanor discourage students from asking questions, if she allows time for questions? Is she unavailable to students outside of class? Does she cancel class frequently? Does she return student work without comment, when she returns it? Does she refuse all invitations to serve on qualifying exam or dissertation committees and to supervise students?

U.C. Berkeley's policy for evaluating teaching (http://teaching.berkeley.edu/campus-and-office-president-policies-evaluating-teaching) advocates looking at such things. It points out many of the limitations of student evaluations of teaching listed above. For instance, it warns, "students should not be used to judge the adequacy, relevance, and timeliness of the course content nor the breadth of the instructor's knowledge and scholarship." Instead, the policy urges departments to develop appropriate procedures that are "supportive and encouraging rather than investigative or punitive" to answer "the essential question … whether the candidate contributes in an effective, creative, and appropriate way to the teaching mission of the department."

The policy advocates using a spectrum of data:

> Combining sources and methods, it is possible to collect a variety of information about a faculty member's teaching. For example, colleagues can evaluate instructional materials or observe an instructor's classroom teaching. Students can complete evaluation forms at the end of a course, participate in individual or group interviews, or fill out surveys when they graduate. …
>
> The candidate's faculty colleagues who have appropriate expertise in the discipline are best able to evaluate the scholarship that informs the design and organization of courses and curriculum, the choice or development of texts and other instructional materials (syllabus, handouts, etc.), the nature of examinations and assignments, and so on.

The Department of Statistics at UC Berkeley has adopted this as standard practice, starting with a pilot in Spring 2013. Every promotion candidate is expected to produce a teaching portfolio for reviews, consisting of a teaching statement, syllabi, notes, websites, assignments, exams, videos, statements on mentoring, or any other materials the candidate feels are relevant. The chair and *ad hoc* committee read and comment on the portfolio in the review. At least before every "milestone" review (mid-career, tenure, full, step VI), a faculty member observs at least one of the candidate's lectures and comments on it, in writing (see http://teaching.berkeley.edu/peer-review-course-instruction for a guide to peer observation of teaching). These in-class observations complement the portfolio and student comments. Distributions of student evaluation scores are reported, along with response rates. Averages of student evaluation scores are downplayed or not reported.

In the pilot, the faculty observer was a member of the Academic Senate's Committee on Teaching who had received the campus's Distinguished Teaching Award. The decision for the observer to be external to the department was not taken lightly. There may be disciplines or departments in which it would be better to have a department member observe. The process included conversations between the candidate and the observer to set the stage, the opportunity for the candidate to respond to the written comments, and a provision for a "no-fault do-over" at the candidate's sole discretion. The candidates and the reviewer reported that the process was valuable and interesting. The faculty observer reported that it took about four hours to communicate with the candidates, schedule the visit, observe the class, and write up his observations. If that is typical, peer observation before four major career milestones would take about 16 hours of effort over someone's career.

Observing more than one class session and more than one course would be better. Adding informal classroom observation—and discussion—between reviews would be better. And periodic surveys of former students, advisees, and graduate teaching assistants would bring another, complementary source of information about teaching. However, using teaching portfolios and even a little classroom observation improves on the *status quo ante*.

In contrast to the example at the beginning of this article, we think an example of a more serious and useful evaluation of teaching might read as follows:

> Smith is, by all accounts, an excellent teacher, as confirmed by the classroom observations of Professor Jones, who calls out Smith's ability to explain key concepts in a broad variety of ways, to hold the attention of the class throughout a 90-minute session, to use both the board and slides effectively, and to engage a large class in discussion. Prof. Jones's peer observation report is included in the case materials; conversations with Jones confirm that the report is his candid opinion: He was impressed, and commented in particular on Smith's rapport with the class, her sensitivity to the mood in the room and whether students were following the presentation, her facility in blending derivations on the board with projected computer simulations to illustrate the mathematics, and her ability to construct alternative explanations and illustrations of difficult concepts when students did not follow the first exposition.
>
> Smith's student teaching evaluation scores are consistently high. Smith's classroom skills are evidenced by student comments in teaching evaluations and by the teaching materials in her portfolio.
>
> Examples of comments on Smith's teaching include:
>
>> I was dreading taking a statistics course, but after this class, I decided to major in statistics.
>>
>> the best I've ever met…hands down best teacher I've had in 10 years of university education
>>
>> overall amazing…she is the best teacher I have ever had
>>
>> absolutely love it
>>
>> loves to teach, humble, always helpful
>>
>> extremely clear … amazing professor
>>
>> awesome, clear
>>
>> highly recommended
>>
>> just an amazing lecturer

>> great teacher … best instructor to date
>>
>> inspiring and an excellent role model
>>
>> the professor is GREAT
>
> Critical student comments primarily concerned the difficulty of the material or the homework. None of the critical comments reflected on the pedagogy or teaching effectiveness, only the workload.
>
> I reviewed Smith's syllabus, assignments, exams, lecture notes, and other materials for Statistics *X* (a prerequisite for many majors), *Y* (a seminar course she developed), *Z* (a graduate course she developed for the revised MA program, which she has spearheaded), and *Q* (a topics course in her research area). They are very high quality and clearly the result of considerable thought and effort.
>
> In particular, Smith devoted an enormous amount of time to developing online materials for *X* over the last five years. The materials required designing and creating a substantial amount of supporting technology, representing at least 500 hours per year of effort to build and maintain. The undertaking is highly creative and advanced the state of the art. Not only are those online materials superb, they are having an impact on pedagogy elsewhere: a Google search shows over 1,200 links to those materials, of which more than half are from other countries. I am quite impressed with the pedagogy, novelty, and functionality. I have a few minor suggestions about the content, which I will discuss with her, but those are a matter of taste, not of correctness.
>
> The materials for *X* and *Y* are extremely polished. Notably, she assigned a term project in an introductory course, harnessing the power of inquiry-based learning. I reviewed a handful of the term projects, which were ambitious and impressive. The materials for *Z* and *Q* are also well organized and interesting, and demand an impressively high level of performance from the students. The materials for *Q* include a great selection of data sets and computational examples that are documented well. Overall, the materials are exemplary; I would estimate that they represent well over 1,500 hours of development during the review period. Conversations with GSIs indicate that Smith spent a considerable amount of time mentoring them, including weekly meetings and observing their classes several times each semester. She also played a leading role in revising the PhD curriculum in the department.
>
> Smith has been quite active as an advisor to graduate students. In addition to serving as a member of sixteen exam committees and more than a dozen MA and PhD committees, she advised three PhD recipients (all of whom got jobs in top-ten departments), co-advised two others, and is currently advising three more. She advised two MA recipients who went to jobs in industry, co-advised another who went to a job in government, advised one who changed advisors. She is currently advising a fifth. She supervised three undergraduate honors theses and two undergraduate internships during the review period.
>
> This is an exceptionally strong record of teaching and mentoring for an assistant professor. Prof. Smith's teaching greatly exceeds expectations.

We feel that a review along these lines would better reflect whether faculty are dedicated teachers, the effort they devote, and the effectiveness their teaching; would comprise a much fairer assessment; and would put more appropriate attention on teaching at any institution.

Recap

- We might wish we could measure teaching effectiveness reliably simply by asking students whether teaching is effective, but it does not work.
- Controlled, randomized experiments—the gold standard for reliable inference about cause and effect—have found that student ratings of teaching effectiveness are negatively associated with direct measures of effectiveness. Student teaching evaluations can be influenced by the gender, ethnicity, and attractiveness of the instructor.
- Summary items such as "overall effectiveness" seem most susceptible to extraneous factors.
- Student comments contain valuable information about students' experiences.
- Survey response rates matter. Low response rates need not signal bad teaching, but they make it impossible to generalize reliably from the respondents to the whole class.
- It is practical and valuable to have faculty observe each other's classes at least once between "milestone" reviews.
- It is practical and valuable to create and review teaching portfolios.
- Teaching is unlikely to improve without serious, regular attention.

Recommendations

1. Drop omnibus items about "overall teaching effectiveness" and "value of the course" from teaching evaluations: They are misleading.
2. Do not average or compare averages of student rating scores: Such averages do not make sense statistically. Instead, report the distribution of scores, along with the number of responders and the response rate.
3. Pay careful attention to student comments—but understand their scope. Students are the authorities on their experiences in class, but typically are not well situated to evaluate pedagogy generally.
4. Use caution extrapolating student evaluations to the entire class. When response rates are low, extrapolation is unreliable.
5. Avoid comparing teaching in courses of different types, levels, sizes, functions, or disciplines.
6. Use teaching portfolios as part of the review process.
7. Use classroom observation as part of milestone reviews.
8. To improve teaching and evaluate teaching fairly and honestly, spend more time observing the teaching and looking at teaching materials.

Notes on contributors

Philip B. Stark is a Professor of Statistics and Chair of the Department of Statistics at UC Berkeley.

Richard Freishtat is a Senior Consultant in the UC Berkeley Center for Teaching and Learning.

References

1. Abrami, P.C., H.M. Marilyn, and F. Raiszadeh. 2001. Business Students' Perceptions of Faculty Evaluations. *The International Journal of Educational Management* 15(1): 12–22.
2. Ambady, N., and R. Rosenthal. 1993. Half a Minute: Predicting Teacher

Evaluations from Thin Slices of Nonverbal Behavior and Physical Attractiveness. *Journal of Personality and Social Psychology* 64(3): 431.

3. Anderson, K., and E.D. Miller. 1997. Gender and Student Evaluations of Teaching. *PS: Political Science and Politics* 30(2): 216-219.
4. Basow, S.A. 1995. Student Evaluations of College Professors: When Gender Matters. *Journal of Educational Psychology* 87(4): 656-665.
5. Beleche, T., D. Fairris, and M. Marks. 2012. Do Course Evaluations Truly Reflect Student Learning? Evidence from an Objectively Graded Post-Test. *Economics of Education Review* 31(5): 709-719.
6. Braga, M., M. Paccagnella, and M. Pellizzari. 2011. Evaluating Students' Evaluations of Professors. *Bank of Italy Temi di Discussione (Working Paper) No*, *825*.
7. Braskamp, L.A., and J.C. Ory. 1994. *Assessing Faculty Work: Enhancing Individual and Institutional Performance.* San Francisco: Jossey-Bass.
8. Carrell, S.E., and J.E. West. 2008. *Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors* (No. w14081). National Bureau of Economic Research.
9. Cashin, W.E. 1990. “Students Do Rate Different Academic Fields Differently.” In *Student Ratings of Instruction: Issues for improving practice*, edited by M. Theall and J. Franklin, 113-121. San Francisco: Jossey-Bass Inc.
10. Cashin, W.E. 1999. Student Ratings of Teaching: Uses and Misuses. In *Current Practices in Evaluating Teaching: A Practical Guide to Improved Faculty Performance and Promotion/Tenure Decisions*, edited by P. Seldin, 25-44. Bolton, MA: Anker.
11. Cashin, W.E., and V.L. Clegg. 1987. *Are Student Ratings of Different Academic Fields Different?* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.
12. Centra, J.A. 1993. *Reflective Faculty Evaluation: Enhancing Teaching and Determining Faculty Effectiveness.* San Francisco: Jossey-Bass.
13. Centra, J.A. 2003. Will Teachers Receive Higher Student Evaluations by Giving Higher Grades and Less Coursework? *Research in Higher Education* 44(5): 495-518.
14. Clayson, D.E. 2009. Student Evaluations of Teaching: Are They Related to What Students Learn? A Meta-Analysis and Review of the Literature. *Journal of Marketing Education* 31(1): 16-30.
15. Cramer, K.M., and L.R. Alexitch. 2000. Student Evaluations of College Professors: Identifying Sources of Bias. *Canadian Journal of Higher Education* 30(2): 143-64.
16. Cranton, P.A., and R.A. Smith. 1986. A New Look at the Effect of Course Characteristics on Student Ratings of Instruction. *American Educational Research Journal* 23(1): 117-128.
17. Davis, Barbara G. 2009. *Tools for Teaching, 2nd Edition*. San Francisco, CA: John Wiley & Sons.
18. Feldman, K.A. 1978. Course Characteristics and College Students' Ratings of Their Teachers: What We Know and What We Don't Know. *Research in Higher Education* 9: 199-242.
19. Feldman, K.A. 1984. Class Size and College Students' Evaluations of Teachers and Courses: A Closer Look. *Research in Higher Education* 21(11): 45-116.
20. Huff, D. 1954. *How To Lie With Statistics*. New York: W.W. Norton.
21. Marsh. H.W. 2007. Students’ Evaluations of University Teaching:

Dimensionality, Reliability, Validity, Potential Biases and Usefulness. In *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, edited by R. P. Perry & J. C. Smart, 319-383. Dordrecht, The Netherlands: Springer.

22. Marsh, H.W., and T. Cooper. 1980. *Prior Subject Interest, Students Evaluations, and Instructional Effectiveness*. Paper presented at the annual meeting of the American Educational Research Association.
23. Marsh, H.W., and M.J. Dunkin. 1992. Students' Evaluations of University Teaching: A Multidimensional Perspective. In *Higher education: Handbook of theory and research – Vol. 8*, edited by J. C. Smart, 143-234. New York: Agathon Press.
24. Marsh, H.W., and L.A. Roche. 1997. Making Students' Evaluations of Teaching Effectiveness Effective. *American Psychologist* 52: 1187-1197.
25. McCullough, B.D., and D. Radson. 2011. Analysing Student Evaluations of Teaching: Comparing Means and Proportions. *Evaluation & Research in Education* 24(3): 183-202.
26. McKeachie, W. J. 1997. Student Ratings: The Validity of Use. *American Psychologist* 52: 1218-1225.
27. Ory, J.C. 2001. Faculty Thoughts and Concerns about Student Ratings. In *Techniques and Strategies for Interpreting Student Evaluations [Special Issue]. New Directions for Teaching and Learning* 87: 3-15.
28. Overall, J.U., and H.W. Marsh. 1980. Students' Evaluations of Instruction: A Longitudinal Study of Their Stability. *Journal of Educational Psychology* 72: 321-325.
29. Seldin, P. 1999. Building Successful Teaching Evaluation Programs. In *Changing Practices in Evaluating Teaching: A Practical Guide to Improved Faculty Performance and Promotion/Tenure Decisions*, edited by P. Seldin, 213-242. Bolton, MA: Anker.
30. Seldin, P., J.E. Miller, and C.A. Seldin. 2010. *The Teaching Portfolio: A Practical Guide to Improved Performance and Promotion/Tenure Decisions (4th Ed.)*. San Francisco, CA: Jossey-Bass.
31. Short, H., R. Boyle, R. Braithwaite, M. Brookes, J. Mustard, and D. Saundage. 2008. A Comparison of Student Evaluation of Teaching with Student Performance. Paper presented at the annual meeting for OZCOTS 2008: Proceedings of the 6th Australian Conference on Teaching Statistics, Australia.
32. Wachtel, H.K. 1998. Student Evaluation of College Teaching Effectiveness: A Brief Review. *Assessment & Evaluation in Higher Education* 23(2): 191-211.
33. Weinberg, B.A., B.M. Fleisher, and M. Hashimoto. 2007. *Evaluating Methods for Evaluating Instruction: The Case of Higher Education (NBER Working Paper No. 12844)*. http://www.nber.org/papers/w12844
34. Worthington, A.C. 2002. The Impact of Student Perceptions and Characteristics on Teaching Evaluations: A Case Study in Finance Education. *Assessment and Evaluation in Higher Education* 27(1): 49-64.
35. University of California, Berkeley, Academic Senate's Committee on Teaching. 1987. *Policy for the Evaluation of Teaching (for Advancement and Promotion)*. Berkeley, CA: Office of Educational Development.