Student Teaching Evaluations
Inaccurate, Demeaning, Misused
September-October 2003 issue of Academe

Administrators love student teaching evaluations. Faculty need to understand the dangers of relying on these flawed instruments.

By Mary Gray and Barbara R. Bergmann

Fifty years ago, students at Harvard University and the University of California, Berkeley, were publishing guides rating teachers and courses. Irreverent and funny, they featured pungent comments: "Trying to understand Professor X's lectures is like slogging uphill through molasses," or "Dr. Y communicated very closely with the blackboard, but I couldn't tell you what he looks like, as he never faced the class." Unfortunately, what originated as a light-hearted dope sheet for the use of students has, at the hands of university and college administrators, turned into an instrument of unwarranted and unjust termination for large numbers of junior faculty and a source of humiliation for many of their senior colleagues.

In the 1970s, schools started requiring faculty to get students to fill out and turn in teaching evaluation forms to the administration. Administrators soon discovered they had a weapon to use against 50 percent of the faculty: they could proclaim that the half of the faculty with below-average scores in each and every department were bad teachers. They have been at it ever since. When administrators say, as they often do, "We won't tenure Professor X or give Professor Y a salary raise because he or she has teaching evaluations that are below average," they are saying, in effect, that "below average" means bad.

We know of one administration that heroically enlarged the proportion of no-good faculty members to 90 percent by declaring that any junior faculty member who failed to achieve scores in the top tenth percentile could not be promoted. But most administrations are content to bad-mouth a mere 50 percent. (If the "average" administrators use is the median, then exactly half of the faculty will be labeled bad. If they use the mean, the proportion labeled bad will probably be slightly above or below half.)

These administrators treat relative position as if it were an absolute measure of merit. They do not allow for the possibility that some departments will have mostly good teachers, in which case some or even all of those with below-average evaluations will be good teachers. They also do not envision departments in which most of the teachers are poor, in which case some or all of those with above-average evaluations may be poor teachers. It is simply incorrect to assume that each department is half and half, or that a whole university is half and half. A faculty member who gets ratings that are well below average is unlikely to be a shining star of teaching, but he or she may be quite good, valuable to the department and the students, and worthy of tenure and a decent salary.

Administrators who would like to achieve a faculty in which everyone is above average should move to Lake Woebegone, the only place where such a thing is possible. In everyplace else, if all those who were below average were fired, the average would simply rise, and about half the previously "good" teachers would then be below the new average, miraculously reborn as "bad" teachers.

One might argue that administrations should give up using relative order, and instead fix on some particular student evaluation score as the borderline between adequate and inadequate teaching. That would make sense if the ratings actually measured teaching effectiveness, but there is evidence that they do not.

Stephen J. Ceci, a professor at Cornell University, devised an experiment to see what might affect student evaluations. He taught a developmental psychology course twice, the first time using his customary style. The second time, he covered the same material and used the same textbook, but made a big effort to be more exuberant, adding hand gestures and varying the pitch of his voice. He characterized the results as "astounding"—his ratings for the second class soared. The students even gave higher ratings to the textbook. But little if any change occurred in the students' performance on exams. Ceci had pleased the students more, but had conveyed the course material no better.

In other studies, lecture content affected student achievement, but had only a negligible impact on student ratings. In other words, the correlation between student achievement and student ratings was low. Should we be willing to define "effectiveness" merely in terms of student satisfaction? In judging colleagues for tenure or raises, why are faculty so willing to trust judgments made by students in areas beyond their competence to judge?

Students give bad evaluations to those whose ac-

cents differ from those of the students, and to those who teach feared and despised required courses, such as statistics for psychology majors. Daniel Hamermesh of the University of Texas found that better-looking teachers get significantly better ratings. Research by Susan Baslow of Lafayette College has revealed that male students gave better ratings to male professors than to female professors, while female students did the opposite. So at least in disciplines where the students are not predominantly of one sex, women will come out on average with about the same ratings as men. But studies by Sheila Bennett and Anne Statham have shown that women have to (and do) spend more effort and time than men on nurturing behavior to get equivalent ratings.

At most, ratings may identify the very best and the very worst teachers, but they are ill designed to make fine distinctions in the vast intermediate range. Moreover, the use of student evaluations against faculty members appears to adversely affect the educational experience of students. In one survey of faculty, 72 percent said that administrative reliance on student evaluations encourages faculty to water down course content. And a careful study at Duke University by statistician Valen Johnson demonstrated that students' expectations of grades influence their ratings of teachers. His finding provides a powerful incentive for faculty to raise grades. Johnson argues that "the ultimate consequences of such manipulation is the degradation of the quality of education in the United States."

Overreliance on student ratings also deters innovation in subject matter and methodology. An untenured faculty member can't risk trying out a new way to teach that might improve student achievement if the faculty member knows that the old method will produce above-average ratings.

If ratings measure only student satisfaction, how does one assess the real effectiveness of a teacher? Among the many other measures available are student performance on exams and assignments, effectiveness in mentoring students, availability of the instructor, the teacher's commitment to curriculum development, involvement of students in the research of the faculty member, and teaching portfolios prepared by faculty. Considering these measures, however, would require the judgment of faculty peers and would take a lot of time and effort. It is easier to settle on a single simple question: is this faculty member above or below average in student ratings?

Other reasons besides convenience influence the way administrations behave. One is the increased attention to "customer satisfaction" that has developed with the move toward the corporate model in higher education and its concomitant diminishing of the role of faculty in university governance. Another reason is that defining "below average" as bad has the effect of reducing the number of faculty who are granted tenure, and reducing the number of raises that are conferred.

Finally, the reliance on evaluations is bad for the health of relations between students and faculty. Jeffrey Stake, a law professor at Indiana University, argues that asking students their opinions undermines the trust and faith they need to place in the teacher. Instead of saying, "Here is a great scholar and teacher; learn from her what you can," the administration of evaluation forms says to students, "We hired these teachers, but we are not sure they can teach or have taught you enough. Please tell us whether we guessed right."

As an entire career can be terminated by not-good-enough evaluations, the procedure of administering the evaluation instruments and getting them turned in forces on the faculty member what Catholics call "an occasion of sin." The administration sets up a system that presents the faculty with a powerful temptation to cheat, and then has to invent demeaning procedures to prevent cheating. The teacher is explicitly forbidden to touch the evaluation sheets after they have been filled out. A student has to be designated to collect and take them to the appropriate office. This procedure tells the students that the teacher is more than likely to be a cheat and a sneak, who will cook the books if given a chance. Both students and teacher pretend not to notice the shaming involved, but it is palpable in such a situation.

For the most part, faculty have allowed this system to evolve with nary a whimper. Those with above-average scores think "I'm all right, Jack." Those with below-average scores are ashamed and feel they have no standing to complain. But this means of judging teaching has no validity and is demeaning to faculty. Those of us who understand this truth have a responsibility to wake up our colleagues on the faculty and the administration to the facts. Perhaps when we have done so, we can move toward getting rid of this inaccurate, misleading, and shaming procedure.

Mary Gray is professor of mathematics and statistics at American University. Barbara Bergmann, a former AAUP president, is distinguished professor of economics emerita at American University.