Student Evaluations: A Critical Review

by Michael Huemer

Informal student evaluations of faculty were started in the 1960's by enterprising college students.[(1)] Since then, their use has spread so that now they are administered in almost all American colleges and universities and are probably the main source of information used for evaluating faculty teaching performance.[(2)] There is an enormous literature on the subject of student evaluations of faculty (SEF).[(3)] The following is a summary of some developments in that literature that should be of special interest to faculty, with particular emphasis on criticisms of SEF that have emerged recently. But I begin with the arguments in favor of the use of SEF.

1. Reliability and Validity of SEF

A test is said to be "reliable" if it tends to give the same result when repeated; this indicates that it must be measuring *something*. A test is said to be "valid" if it is measuring what it is intended to measure. E.g., a scale that always reads "5" whenever a *red* object is placed on it is "reliable" but not "valid" as a measure of weight.

Most researchers agree

(1) that SEF are highly reliable, in that students tend to agree with each other in their ratings of an instructor, and

(2) that they are at least moderately valid, in that student ratings of course quality correlate positively with other measures of teaching effectiveness. In one type of study, multiple sections of the same course are taught by different instructors, but there is a common final exam. The ratings instructors receive turn out to be positively correlated with the performance of their students on the exam. The correlation is in the neighborhood of .4 to .5, meaning that 16 to 25% of the variance in one variable can be explained by variance in the other.

SEF also tend to correlate well with retrospective evaluations by alumni; in other words, former students rarely change their evaluations of their teachers as the years pass.[(4)]

Furthermore, other methods of evaluating teaching effectiveness do not appear to be valid. Ratings by colleagues and trained observers are not even reliable (a necessary condition for validity)--that is, colleagues and observers do not even substantially agree with each other in instructor ratings.[(5)]

2. Usefulness of SEF

Instructors who received results of a midsemester evaluation tended to have higher ratings on end-of-semester evaluations than those who did not, suggesting that SEF cause changes in teaching behaviors which result in higher ratings. The improvement was greatest when (a) the professor's self-evaluation was very different from the students' evaluation, (b) the professor received professional consultation on the interpretation of the evaluations, and (c) the student evaluation forms included specific items (such as, "Professor gives preliminary overview of lecture"), as opposed to vague items such as, "How well planned are lessons?"

In spite of the above, SEF have come under fire on several fronts.

3. Grading Leniency Bias

The most common criticism of SEF seems to be that SEF are biased, in that students tend to give higher ratings when they expect higher grades in the course. This correlation is well-established, and is of comparable magnitude, perhaps larger, to the magnitude of the correlation between student ratings and student learning (as measured by tests) described in section 1 above. Thus, SEF seem to be as much a measure of an instructor's leniency in grading as they are of teaching effectiveness. The correlation holds both between students in a given class and between classes. It also holds between classes taught by the same instructor, when the instructor varies the grade distribution. And it affects ratings of all aspects of the instructor and the course.[(6)] Many believe that this causes rampant grade inflation.[(7)]

Optimists have suggested that this correlation might be due to the fact that greater teaching effectiveness on the part of the instructor leads to both higher grades and higher ratings of the instructor; thus, the effect might actually be a sign of the validity of student ratings. However, this hypothesis fails to explain (a) why the correlation also holds among students within the same class (who presumably are beneficiaries of the same teaching effectiveness), (b) why it holds between classes taught by the same instructor when the instructor varies the grade distribution, (c) why there is a greater correlation between grades and ratings when one looks at the student's *relative* grade (i.e., the student's grade in this class compared with his/her grade in other classes), as opposed to the student's absolute grade. These and other facts are explained by the leniency bias hypothesis: people tend to like those who praise them (particularly if the praise is greater than expected) and dislike those who criticize them. The instructor who grades leniently in effect praises the students, who then like the instructor more. They then reward the instructor with higher ratings in general.[(8)]

Despite some dissenting voices,[(9)] the influence of grades on student evaluations seems to be an open secret in colleges and universities. In one survey, 70% of students admitted that their rating of an instructor was influenced by the grade they expected to get.[(10)] Similar proportions of professors believe that grading leniency and course difficulty bias student ratings.[(11)]

4. Dumbing Down Courses

A related complaint many have is that SEF encourage professors to dumb down courses in an effort to keep students happy at all costs. In one survey, 38% of professors admitted to making their courses easier in response to SEF.[(12)]

Peter Sacks provides a more detailed, though anecdotal picture. Sacks reports having almost lost his job due to low teaching evaluations from his students. He was able to dramatically raise his teaching evaluations and gain tenure, he says, by becoming utterly undemanding and uncritical of his students, giving out easy grades, and teaching to the lowest common denominator. Sacks claims that this behavior is not unusual but is rather the norm at his college, where students are king and entertainment is all that matters. An excerpt from Sacks' book:

And so, in my mind, I became a teaching teddy bear. In the metaphorical sandbox I created, students could do no wrong, and I did almost anything possible to keep all of them happy, all of the time, no matter how childish or rude their behavior, no matter how poorly they performed in the course, no matter how little effort they gave. If they wanted their hands held, I would hold them. If they wanted a stapler (or a Kleenex) and I didn't have one, I'd apologize. If they wanted to read the newspaper while I was addressing the class or if they wanted to get up and leave in the middle of a lecture, go for it. Call me spineless. I confess. But in the excessively accommodative culture that I found myself in, "our students" as many of my colleagues called them, had too much power for me to afford irritating them with demands and challenges I had previously thought were part and parcel of the collegiate experience.[(13)]

5. Educational Seduction, or the Dr. Fox Effect

In a well-known study, a professional actor was hired to deliver a non-substantive and contradictory lecture, but in an enthusiastic and authoritative style. The audience, consisting of professional educators, had been told they would be listening to Dr. Myron Fox, an expert on the application of mathematics to human behavior. They were then asked to rate the lecture. Dr. Fox received highly positive ratings, and no one saw through the hoax.[(14)] Later studies have obtained similar results,[(15)] showing that audience ratings of a lecture are more strongly influenced by superficial stylistic matters than by content.

Adding support to this conclusion was another study, in which students were asked to rate instructors on a number of personality traits (e.g., "confident," "dominant," "optimistic," etc.), on the basis of 30-second video clips, without audio, of the instructors lecturing. These ratings were found to be very good predictors of end-of-semester evaluations given by the instructors' actual students. A composite of the personality trait ratings correlated .76 with end-of-term course evaluations; ratings of instructors' "optimism" showed an impressive .84 correlation with end-of-term course evaluations. Thus, in order to predict with fair accuracy the ratings an instructor would get, it was not necessary to know anything of what the instructor said in class, the material the course covered, the readings, the assignments, the tests, etc.[(16)]

Williams and Ceci conducted a related experiment. Professor Ceci, a veteran teacher of the Developmental Psychology course at Cornell, gave the course consecutively in both fall and spring semesters one year. In between the two semesters, he visited a media consultant for lessons on improving presentation style. Specifically, Professor Ceci was trained to modulate his tone of voice more and to use more hand gestures while speaking. He then proceeded, in the spring semester, to give almost the identical course (verified by checking recordings of his lectures from the fall), with the sole significant difference being the addition of hand gestures and variations in tone of voice (grading policy, textbook, office hours, tests, and even the basic demographic profile of the class remained the same). The result: student ratings for the spring semester were far higher, usually by more than one standard deviation, on all aspects of the course and the instructor. Even the textbook was rated higher by almost a full point on a scale from 1 to 5. Students in the spring semester believed they had learned far more (this rating increased from 2.93 to 4.05), even though, according to Ceci, they had not in fact learned any more, as measured by their test scores. Again, the conclusion seems to be that student ratings are heavily influenced by cosmetic factors that have no effect on student learning.

6. Academic Freedom

Some argue that SEF are a threat to academic freedom.[(17)] Not only do SEF influence instructors' grading policies, teaching style, and course difficulty, but they may also restrict what a professor says in class. Professors may feel inhibited from discussing controversial ideas or challenging students' beliefs, for fear that some students will express their disagreement through the course evaluation form. More than one author has described SEF as "opinion polls," with the suggestion that SEF require professors to think like politicians, seeking to avoid giving offense and putting style before substance.[(18)]

Alan Dershowitz reports that some of his students have "used the power of their evaluations in an attempt to exact their political revenge for my politically incorrect teaching." One student, who complained to Dershowitz about his (Dershowitz') teaching about rape from a civil liberties perspective, informed Dershowitz that he should expect to be "savaged" on the student evaluations at the end of the term. Several students subsequently complained on their teaching evaluations about the content of his lectures on the subject of rape, saying that they were offensive, that he should not be allowed to teach at Harvard, and so on. Alan Dershowitz, of course, need have little fear of losing his job. The same is not true of less prominent, junior faculty at institutions across the country.[(19)] I have personally received evaluation forms

complaining that the professor "teaches his own views," and I have as a result been influenced to remove controversial material from my classes.

College students do not have a universal appreciation for the ideals of free speech and academic freedom. An anthropology professor I once had at Berkeley became locally (in)famous for his criticisms of affirmative action and for his view that minorities and women had lower average levels of intelligence than the rest of the population. Subsequently, a group of students disrupted his class to protest against his allegedly racist, sexist, and homophobic teachings. The students went on to call for his dismissal from the university. Signs appeared on campus saying, "No more racist bullshit in the name of academic freedom."[(20)] Berkeley, it seemed, had come a long way since the days of the free speech movement. Fortunately for him, the professor already had tenure. But what would have happened to a junior faculty member in a similar position? Given the student reaction in this case, it is not difficult to imagine that even much less controversial statements might have elicited low end-of-term evaluations from those students who wished to see the professor fired. Even a small percentage of such extremely negative evaluations could have a significant impact on a professor's career.

Professors discussing unconventional or controversial ideas *may* also receive a larger number of very positive student evaluations, relative to other professors whose classes are more bland and, perhaps, boring. In spite of this, there are two reasons why the overall incentive created by SEF will be for the professor to avoid controversy. First, the average rating professors receive is 4 or above on a scale of 1 - 5; therefore, a very hostile student can give a rating three points below the average, whereas a very enthusiastic student can only give a rating one point above the average. Thus, assuming the professor is average, the marginal unusually hostile student has an impact up to three times greater than the marginal unusually enthusiastic student. Second, there is a saying in American politics to the effect that one doesn't gain votes, one only loses them--meaning that it is much easier to earn a voter's opposition through taking substantive stands on issues than it is to gain support by doing so. If a politician says three things that I agree with and one that I disagree with (all concerning emotionally charged issues), I am more likely to vote against him, provided the other candidate did not say anything I disagreed with, even if this was because the latter said very little at all. This explains why American politicians often avoid taking non-trivial stands on issues. A similar principle applies to professors, when their retention is decided in a similar manner: any statement or question a teacher raises that anyone could take offense at will run a risk of evoking hostile reactions from a few students who will regard the statement or question as grounds for a negative evaluation, while there is little chance that even a non-hostile student will take it as grounds for an especially positive evaluation. Thus, it is reasonable to suppose that the degree to which a professor is controversial would be a strong depressive factor on his student evaluations, although this thesis has not yet been subjected to systematic testing.

There exist simple and well-known ways for a professor to avoid giving offense. One technique, when a class ostensibly focuses on a controversial subject matter, is to focus one's lectures on what other people have said. For example, a professor may, without raising any eyebrows, teach an entire course of lectures on ethics without ever making an ethical statement, since he confines himself to making reports of what other people have said about ethics. This ensures that no one can take offense towards *him*. During classroom discussions, he may simply nod and make non-committal remarks such as "Interesting" and "What do the rest of you think about that?", regardless of what the students say. (This provides the added "advantage" of reducing the need both for preparation before class and for effort during class, on the part of the professor.) Although pedagogic goals may often require correcting students or challenging their logic, SEF-based performance evaluations provide no incentive to do so, while the risk of reducing student happiness provides a strong incentive not to do so. Some students may take offense, or merely experience negative feelings, upon being corrected, whereas it is unlikely that students would experience such negative feelings as a result of a professor's failure to correct them. Overall, SEF reward professors who tell their students what they want to hear.

G. F. Schueler draws our attention to a related case:

Socrates, who is usually thought to have been one of the world's "Great Teachers," nevertheless received rather low marks from his "students" on his final teaching evaluation. At a time of life when most of us would long since have retired, the Athenian jurors at his trial met his request for a pension by voting to put him to death...[(21)]

As Schueler notes, there is no reason to believe that the majority of Athenian citizens who were familiar with Socrates' activities would have evaluated his work as a philosopher much differently. The death sentence, allegedly for corrupting the youth and believing in gods of his own invention, was Socrates' payment for his lifelong efforts at challenging the beliefs of his fellow citizens. Though today's students lack the power to put to death professors with whom they disagree, the lesson that such challenges are not always welcome is unlikely to be lost on those professors who hold unconventional views.

7. Why Use SEF?

In the light of the preceding objections, why do most institutions continue to use SEF? The main reasons are probably the following: (a) SEF are easy and inexpensive to administer. (b) SEF give an impression of objectivity, in comparison with more "subjective" measures such as letters from observers, since SEF produce definite *numbers*. (The impression seems to be an illusion, however, since the numbers are merely measurements of subjective impressions.) (c) There are few alternatives to SEF, if one wants to assess teaching effectiveness. Steven Cahn argues that teaching should be assessed by experts in the field, i.e., one's colleagues,[(22)] but as indicated in section 1, such measures appear to be even less valid. Greenwald and Gillmore suggest using student ratings but with statistical corrections for grading leniency; this, however, would not address the concerns of sections 4, 5, and 6 above.

8. Other Approaches

Institutions seeking to improve teaching quality may take one or more of the following measures, which would not be subject, or would be less subject, to the objections of sections 3-6:

1. Faculty members could be offered courses or workshops on improving teaching effectiveness, receiving recognition on performance reviews for having taken such courses.

2. Student evaluation forms could be redesigned to emphasize relatively objective matters, such as "Did the professor come to class on time?", "Did he read student work and return it within a reasonable time frame?", and so on, rather than subjective items such as "How would you rate this instructor?" or "How fair was the grading?" The former sort of questions would probably be less subject to the effects of bias than the latter. In addition, they have a better chance of inducing improvements in teaching performance.

3. Written comments might be taken into account in weighting student ratings. Evaluation forms on which low ratings are given without explanation, or where the complaints are directed at the professor's beliefs, the harshness of the grading, the difficulty of the course, or the professor's personal characteristics (such as physical appearance, clothing style, or personality) might be discounted.

4. Teaching can be evaluated in part by examination of syllabi and other course materials. These can be used to verify that a course contains substantive content; but professors should not be monitored for the "correctness" or moral or political value of that content.

9. The Philosophy of Consumerism

A fourth reason why SEF are widely used may be the belief that the university is a business and that the responsibility of any business is to satisfy the customer. Whether they measure teaching effectiveness or not, SEF are probably a highly accurate measure of student satisfaction (and the customer is always right, isn't he?). However, even if we agree to view the university as a business, the preceding line of thought rests upon a confusion about the product the university provides. Regardless of what they may themselves think at times, students do not come to college for entertainment; if they did, they might just as well watch MTV for four years and put that on their resumes. Students come to college for a diploma. A diploma is a certification by the institution that one has completed a course of study and thereby been college-educated. But that will mean nothing unless the college or university can maintain intellectual standards. A particular student may be happy to receive an easy A without having to work or learn much, but a college that makes a policy of providing such a product will find its diplomas decreasing in value.

Part of a university's responsibility may be to satisfy its students. But it is also a university's responsibility to educate those individuals whom it is certifying as educated. Unfortunately, those goals are often in conflict.

References

Abrami, Philip C., Les Levanthal, and Raymond P. Perry. "Educational Seduction," *Review of Educational Research* 52 (1982): 446-64.

Ambady, Nalini and Robert Rosenthal. "Half a Minute: Predicting Teacher Evaluations from Thin Slices of Nonverbal Behavior and Physical Attractiveness," *Journal of Personality and Social Psychology* 64 (1993): 431-41.

Cahn, Steven M. *Saints and Scamps: Ethics in Academia* (Totowa, NJ: Rowman & Littlefield, 1986).

Cave, Martin, Stephen Hanney, Mary Henkel, and Maurice Kogan. *The Use of Performance Indicators in Higher Education: The Challenge of the Quality Movement*, 3rd ed. (London: Jessica Kingsley Publishers, 1997).

Centra, John A. *Reflective Faculty Evaluation* (San Francisco: Jossey-Bass Publishers, 1993).

d'Apollonia, Sylvia and Philip C. Abrami. "Navigating Student Ratings of Instruction," *American Psychologist* 52 (1997): 1198-1208.

Dershowitz, Alan. *Contrary to Popular Opinion* (New York: Pharos Books, 1992).

Gilbaugh, John W. "Renner Substantiated," *Phi Delta Kappan* 63 (Feb. 1982): 428.

Goldman, Louis. "The Betrayal of the Gatekeepers: Grade Inflation," *Journal of General Education* 37 (1985): 97-121.

Greenwald, Anthony G. and Gerald M. Gillmore. "Grading Leniency Is a Removable Contaminant of Student Ratings," *American Psychologist* 11 (1997): 1209-17.

Haskell, Robert E. "Academic Freedom, Tenure, and Student Evaluation of Faculty: Galloping Polls in the 21st Century," *Education Policy Analysis Archives* 5 (1997). Available online at <http://olam.ed.asu.edu/epaa.v5n6.html>.

Marsh, Herbert W. "Student Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research," *International Journal of Educational Research* 11 (1987): 253-388.

Marsh, Herbert W. and Lawrence A. Roche. "Making Students' Evaluations of Teaching Effectiveness Effective," *American Psychologist* 52 (1997): 1187-97.

Naftulin, Donald H., John E. Ware, and Frank A. Donnelly, "The Doctor Fox Lecture: A Paradigm of Educational Seduction," *Journal of Medical Education* 48 (1973): 630-5.

Rice, Lee. "Student Evaluation of Teaching: Problems and Prospects," *Teaching Philosophy* 11 (1988): 329-44.

Ryan, James J., James A. Anderson, and Allen B. Birchler, "Student Evaluations: The Faculty Responds," *Research in Higher Education* 12 (December, 1980): 317-33.

Sacks, Peter. *Generation X Goes to College* (LaSalle, IL: Open Court, 1986).

Schueler, G. F. "The Evaluation of Teaching in Philosophy," *Teaching Philosophy* 11 (1988): 345-8.

Selvin, Paul. "The Raging Bull of Berkeley," *Science* 251 (1991): 368-71.

Williams, Wendy M. and Stephen J. Ceci. "'How'm I Doing?' Problems with Student Ratings of Instructors and Courses," *Change: The Magazine of Higher Learning* 29 (Sept./Oct. 1997): 12-23.

Wilson, Robin. "New Research Casts Doubt on Value of Student Evaluations of Professors," *Chronicle of Higher Education* (Jan. 16, 1998): A12.

Notes

1. Cahn, 37.

2. Cave, et al., 147; Haskell; d'Apollonia and Abrami, 1198; Wilson.

3. According to Wilson, nearly 2000 studies of SEF have been completed.

4. For a summary of the data on reliability and validity, see Centra, 58-65.

5. Marsh and Roche, 1190.

6. See Rice, 335-6; Wilson; Greenwald and Gillmore, 1214.

7. See Goldman; Sacks.

8. See Greenwald and Gillmore. The authors discuss five alternative interpretations of the grades-ratings correlation, arguing that only the leniency-bias hypothesis explains all the patterns in the data.

9. d'Apollonia and Abrami, 1204-5.

10. See Gilbaugh, who reports that 360 of 518 students surveyed at San Jose State University gave the response indicated. This result may be taken with a grain of salt, as Gilbaugh reports it in a letter to the editor and does not give details as to survey methods. However, the results are more likely an underestimate than an overestimate, both because students may be reluctant to admit to what most would regard as unfair behavior on their part and because some students may be unaware of their bias.

11. See Marsh.

12. Ryan et al.

13. Sacks, 85.

14. Naftulin, et al.

15. See Abrami, Leventhal and Perry. However, the authors caution that these results provide little information about the validity of student ratings, in part because it is not known how much either content or stylistic factors vary among actual college professors. If, for instance, actual professors varied very little in presentation styles, then the Dr. Fox effect would not be relevant in most cases.

16. See Ambady and Rosenthal.

17. See Haskell.

18. Williams and Ceci, 12, 23; Schueler.

19. Dershowitz, 117-19.

20. The incident is discussed in Selvin.

21. Schueler, 345.

22. Cahn, 36-41.