[Published originally in the May 2003 edition of *Computing Research News*, Vol. **15**/No. 3, pp. 2, 10.]

Expanding the Pipeline

Are Student Evaluations of Teaching Fair?

By Faith E. Fich[1]

Introduction

Anonymous student evaluations of teaching are widely used at universities throughout North America for tenure and promotion decisions, determination of yearly salary increases, and the choice of teaching award recipients. Their purpose is to fairly evaluate the teaching quality of faculty members and help them improve their teaching. Yet, the perception of many faculty members, including me, is that the use of student evaluations of teaching achieves neither of these goals.

Over the past few years, I have talked to many faculty members in science and engineering about student evaluations, including department chairs, deans, and directors of a number of university teaching centers. I was also a member of a committee set up by the dean of the Faculty of Arts and Science at the University of Toronto to address the issue of bias in evaluation of women science professors' teaching.

It is outside the scope of this article to give a survey of the relevant literature. However, I'll mention some of the evidence that has convinced me of the unfairness of student evaluations of teaching as they are often used. More importantly, I will discuss implications for how the results of student evaluations should be used and present a few general recommendations for evaluating teaching fairly and effectively.

Gender Bias

I want to begin with one carefully controlled experiment that I found very compelling. Sinclair and Kunda [6] administered a test of 10 open-ended questions to approximately 50 male students. Each student was given feedback on his performance, randomly chosen from among four prerecorded videos. There were two evaluators, one male and one female, and two scripts, one praising the student's performance and one criticizing it. After receiving their feedback, each student was asked to rate his evaluator.

The results of this experiment are summarized by the title of their paper, "Motivated Stereotyping of Women: She's Fine if She Praised Me but Incompetent if She Criticized Me."

[1] Faith E. Fich (fich@cs.toronto.edu) is a Professor in the Department of Computer Science at the University of Toronto.

More precisely, among ratings by students who had been given positive feedback, the two evaluators were rated roughly the same. However, among ratings by students who had been given negative feedback, the female evaluator was rated significantly lower than the male evaluator.

To isolate the cause of the difference in ratings, the experiment had a second part. In it, the test answers given by each student and the evaluation he had been given were shown to an observer, another male student who had not taken the test. Each observer was also asked to rate the evaluator. Among the observers, the ratings that evaluators received were not correlated with their gender.

Sinclair and Kunda's interpretation is that when people are criticized, they unconsciously use negative stereotypes about the criticizer to discount the validity of the criticism, as a way of maintaining self-esteem. An implication is that women professors who have high standards or who teach courses that students find difficult may well be victims of bias. They obtained similar results studying racial bias [5].

Another interesting experiment was performed by Kaschak [3]. A set of 25 male students and 25 female students were asked to rate professors, given descriptions of the professors and their teaching methods. Half the professors were listed as female, the other half as male. A second set of 25 male students and 25 female students were given the same descriptions, with the genders of the professors switched. Although the gender of the professor did not affect the ratings by female students, the male students rated the female professors lower.

Other Factors Affecting Student Evaluations

There is a vast body of literature about student evaluations of teaching, containing many conflicting conclusions. The problem is that there are many variables unrelated to the quality of teaching that may affect evaluations and that interact in complex ways. Furthermore, most of this work consists of statistical analyses, where factors that are significant for a small segment of the population, for example, women computer science professors, can be insignificant in the aggregate data.

Nevertheless, the bulk of the research does show that certain factors unrelated to teaching quality do affect students' evaluations of teaching. Students in higher-level courses tend to rate professors more favorably than students in lower-level courses [1, 4]. The same is true for students taking elective courses as compared with students taking required courses [1, 4]. Both of these outcomes may be related to the students' greater interest in the course material. There is also evidence that small class size [4] and leniency of grading [2, 4] lead to better ratings.

Faculty who penalize students for committing plagiarism may receive unfairly low ratings from those students. In computer science courses, it is relatively easy for students to copy pieces of code from one another and there is sophisticated software to detect plagiarism. Thus, this factor may have a greater effect on our evaluations than in other disciplines.

How to Use Student Evaluations

In light of the many factors unrelated to the quality of teaching that can affect the results, it is important to recognize the limitations of student evaluations of teaching when they are used.

Only compare results from similar courses.

In particular, comparisons of results should only occur for faculty members teaching courses with similar characteristics. These include factors such as class size, course level, difficulty of the material, whether the course is theoretical or applied, and whether the course is required.

Avoid general subjective items.

Bias is more likely to affect general subjective items such as "overall effectiveness of instructor." Therefore, such items should be avoided, especially for tenure, promotion, or salary considerations [1], in spite of the fact that some administrators want a single summary number.

Be careful with new courses.

When a course is taught for the first time or in a significantly different way, students' experiences often do not match their prior expectations. This can cause unfairly negative ratings, and the results must be interpreted with care. In particular, they should not be used to justify an unsuccessful tenure or promotion decision.

Results of student evaluations have low precision.

There are known problems with the precision of the results of student evaluations of teaching. For example, in my department, there are some multi-section courses where there are common assignments graded by the same teaching assistant. Yet, there are often significant differences in the student responses to supposedly objective items such as "assignments are graded fairly" or "returns work promptly," for faculty who are teaching different sections. These results seem to be correlated with the responses to general subjective items. In science, high-precision conclusions require justification. Given the large number of possible sources of error that can arise when students evaluate teaching, it is only reasonable to interpret the results using a very coarse scale: outstanding, good, or poor.

The teaching evaluation forms used by the Faculty of Science at McMaster University in Canada address this problem by giving students a comprehensive list of possible positive and negative characteristics about a course and the people who teach it. Students are asked to choose particularly relevant items from this list, rather than rating a small number of aspects on a scale.

Eliminate inappropriate student comments.

Some students use anonymous student evaluations to make inappropriate, slanderous, or abusive comments. For example, one woman received the comment, "She should wear more provocative clothing." The number of such comments seems to be significantly higher in biology and computer science departments, where there are many students who are not motivated by interest in learning or the subject matter in their courses, but rather by hopes of admission to medical school or high-paying jobs in IT.

University staff administering the evaluations should remove forms containing biased or offensive comments and not include their ratings in any compilations. However, the number of forms containing such comments should be reported as possible evidence of bias. A better way to discourage inappropriate comments is to partly remove students' anonymity: Enable staff to identify the student evaluations, but don't allow any faculty members (including department chairs) to access this information. Texas A&M University has used this approach for many years. Electronic evaluations are a good way to implement partial anonymity. In addition, they have the advantage of giving students ample time to express their views, and they do not necessarily exclude students who missed class on a particular day.

General Recommendations

Use multiple forms of evaluation.

Most teaching experts agree that multiple forms of evaluation are needed to properly evaluate teaching. One reason is that many aspects of teaching cannot be addressed by students. Alternative methods of evaluation are especially important for professors who receive (possibly unjustified) low ratings from students or when there is a perception of possible bias (such as high standard deviation in student ratings, inconsistent ratings in different classes, or inappropriate comments, anonymous notes, or newsgroup postings).

Other ways to evaluate teaching include observation of lectures and examination of course material by trained peers or teaching experts (e.g., from a university's teaching center), the use of teaching portfolios, giving exit interviews or questionnaires to graduating students, requesting letters from former students, having student discussion facilitated by trained faculty or staff, obtaining feedback from teaching assistants, and comparing the performance of students on common, jointly-graded exams in multi-section courses taught by different professors. Peer and expert evaluation can be particularly valuable in helping professors improve their teaching because the criticism is likely to be constructive and objective. Furthermore, professors are generally more receptive to their feedback.

More detailed assessment should be done periodically.

Because it is more expensive, detailed assessment of teaching might not be done every year. However, it should be done periodically, say once every five years and more frequently prior to tenure. Courses, themselves, should also be evaluated. The curriculum may require a course to cover too much material, or the background, ability, or motivation of students enrolling in the course may have changed. In such situations, it is unfair to penalize a teacher who is attempting to meet unrealistic requirements.

Have a transparent teaching evaluation process.

It is important that detailed written information be provided to faculty outlining explicit expectations for good teaching and explaining how teaching evaluations affect salary, tenure, and promotion decisions. This should include what information is considered, what criteria are used, whether comparisons are being made, and, if so, with whom and why. When there is a possibility of bias in some of the information, this fact and how it is dealt with should be mentioned. If improvement in teaching is needed, specific objectives and ways of achieving those objectives should be discussed with that faculty member.

Why should departments care about improving their teaching evaluation process? If it has been done the same way for a long time and there haven't been major problems, why should it change? One reason is that even a slightly biased process can, over time, lead to substantial inequities in salary. Another reason is that when faculty members have the perception that they are being unfairly evaluated, they feel unappreciated. This can affect their morale, the effort they are willing to put towards teaching, and their desire to stay in their department. Thus, improving the teaching evaluation process might improve retention as well as teaching.

References

[1] Raoul Arreola, *Developing a Comprehensive Faculty Evaluation System*, Anker Publishing Co., 1995.

[2] Anthony Greenwald and Gerald Gillmore, Grading Leniency Is a Removable Contaminant of Student Ratings, *American Psychologist*, **52**(11), 1997, pp. 1209-1217.

[3] Ellyn Kaschak, Sex Bias in Student Evaluations of College Professors, *Psychology of Women Quarterly*, **2**(3), 1978, pp. 235-242.

[4] Ian Neath, How to Improve Your Teaching Evaluations Without Improving Your Teaching, *Psychological Reports*, **78**, 1996, pp. 1363-1372.

[5] Lisa Sinclair and Ziva Kunda, Reactions to a Black Professional: Motivated Inhibition and Activation of Conflicting Stereotypes, *Journal of Personality and Social Psychology*, **77**(5), 1999, pp. 885-904.

[6] Lisa Sinclair and Ziva Kunda, Motivated Stereotyping of Women: She's Fine if She Praised Me but Incompetent if She Criticized Me, *Personality and Social Psychology Bulletin*, **25**(11), 2000, pp. 1329-1342.