Why the University Should Abolish Faculty Course Evaluations

Clark Glymour
Draft, November, 2003

In company with several other University bodies, the Faculty Senate has recently approved a new Faculty Course Evaluation instrument, to be filled out on-line. Both for students and for faculty the decision is unfortunate, and quite possibly for some faculty, sometime, it will be very unfortunate. What the Faculty Senate ought to have done was to recommend investigation of a more serious process for estimating the quality of instruction, leading towards the end of University sponsorship of Faculty Course Evaluation forms. Students should be entirely free to organize and publicize their own on-line evaluations of courses and faculty, but the results should not have the *imprimateur* of the University itself. I will start my argument with some anecdotes.

In 1969 Princeton University for the first time admitted about 20 African American students, and nearly half of them enrolled, with sixty other students, in my Introduction to Mathematical Logic. Every one of these smart, brave, ambitious black students failed my course. I thought hard over the summer about why, and formed this hypothesis: the lecture course had based grades on a mid-term and a final and homework. If the African American students were missing some background, or not good at test taking or at judging how well they understood the material, the course structure offered no way for them to make up for those disadvantages by extra effort. The next year I changed the structure of the course. Using a text that divided the material into a great many short chapters with many problems, a student could take a test on a chapter at any time during regular work hours and have it graded immediately; if the test was passed, the student could go on to the next chapter; if the student failed, another test on the same chapter could be taken after a two-day wait; tests on a chapter could be taken repeatedly until one was passed. Lectures were replaced by scheduled problem solving sessions in which the students asked me how to do problems in the text, and I worked the problems out for them; mini-lectures were given spontaneously when students asked about particular material. In addition, I met privately in my office with every enrolled student every other week. Grades were based entirely on how many chapters were successfully completed—how many tests were failed did not matter. The results were interesting: A students mastered almost twice the material I had presented in the previous year; B students somewhat more than the previous year. (For those readers to whom it matters: A students completed prepositional proof theory, semantics and completeness theorem; S5 modal logic proof theory, semantics and completeness theorem; first-order quantification theory rules, semantics and soundness proof.) I had about the same number of African American students as the previous year. Every one of them passed the course with a grade of C or better, and half of them received A grades. Having written as many as 15 exams for each of about 30 chapters, and spent many hours each week meeting with students and graders and reviewing student progress, I was exhausted but exhilarated. Then the FCEs came back, the lowest I have ever received. The student consensus was

that I had contrived the arrangement to save the trouble of preparing lectures. Moral*: Student evaluations are more influenced by formats meeting their expectations than by how much they and their classmates learned.*

Several years ago I served on a committee established by the Dean of Arts and Sciences at the University of Pittsburgh to review the case for tenuring a young mathematics professor there. The man in question had been promoted to Associate Professor without tenure, an entirely anomalous position at Pitt, as at many other universities (but not, of course, at Carnegie Mellon.) The Chair of the Mathematics department made the case for promotion to the committee: the fellow had not been given tenure previously because, although his scholarship was excellent, his faculty course evaluations were unacceptably low, but they had since improved, and so he should now be tenured. Committee heads nodded as the Chair went on about how the Mathematics department valued teaching. I asked the Chair these questions, with the following answers: Was there any evidence other than FCEs that the fellow was a poor teacher? There was not. Prior to the previous decision to promote him without tenure, what had the professor been assigned to teach? Sections of Calculus and of Differential Equations. Were there many such sections? There were. Did they use the same texts and give the same examinations? They did. On average, how did students in his sections do on the final examination compared with students in other sections of the same courses? Here was the give-away*:* On average, they did better than students in other sections. Morals: *Faculty Course Evaluations have little to do with learning, and they can seriously, and unjustly, affect careers.*

From 1984 until 1989 I was Head of the new Philosophy Department at Carnegie Mellon. A newly hired assistant professor consistently received the lowest Faculty Course Evaluations in the department, and I was concerned for his career. I knew the man and his outstanding scholarly work well, and I could guess the problems. He was not charming or funny or good looking, and he had a deep and formal view of philosophical topics, and in his classes he tended to emphasize logical structures and problems imbedded in traditional philosophy. I met with him and urged him to go to the Teaching Center to get advice on how to improve student responses to his teaching. He refused point blank, but assured me his evaluations would improve dramatically. They did. The next semester he had the highest overall course evaluations in the department, and naturally I asked him how he did it—had he changed how he taught, or what he taught? "Not at all," he said, "before the evaluations were given out almost all of the students knew they were going to get A's. I see no reason to sacrifice my career to the cause of grade deflation." Moral*: Faculty course evaluations are substantially influenced by the grades students expect to receive. Basing promotions even in part on faculty course evaluations invites grade inflation and creates an incentive to pander*.

For one year during my headship, John Modell was Acting Dean of the College of Humanities and Social Sciences. He was concerned about Faculty Course Evaluations in the College, and sent around to the various Heads a ranking of the average FCE scores for each department in the College for the previous couple of years. Statistics was at the bottom of the ranking, History, as I recall, at the top. I took the list of departments, but not the rankings, to members of several departments and asked them to rank order the

departments by the amount of mathematical content they believed to be typical of courses in each department. In every case, the rank ordering by mathematical content was the same as Dean Modell's ranking by FCEs. Moral: *Unless students are committed to a mathematical curriculum (as are, for example, majors in Mathematics, Computer Science, Statistics and Physics) the more mathematical content a course has, the less students tend to like it.*

For my first five years at Carnegie Mellon, each semester I taught a required Introductory Philosophy course to about 250 students in one big lecture class. I was interested in whether the course improved students' reasoning abilities, which had not been directly addressed in any way in the course. I asked Jay Devine, Director of Advising for H&SS, if there was some principle by which students were enrolled in the course in Fall rather than Spring semesters, for example, if some judgment was made about students' readiness for the course. He said there was not, that enrollments were driven by scheduling convenience. So I thought students in the course at the end of the Fall semester and at the beginning of the Winter semester were probably roughly comparable groups. I selected a number of general reasoning questions from the Graduate Record Examination and from the Law School Admission Test, and scattered them through the final examination in the Fall term. I gave the same questions scattered through a first examination, before mid-term, in the next Winter semester. Despite the obvious limitations of the comparison, to my pleasure the Fall term students scored about 20% better than the Winter term students. I reported the results to the Dean, who took no interest whatsoever. He did send me a letter congratulating me for high FCE ratings.

One year, instead of giving FCEs, my colleague, Richard Scheines, did a careful study in an introductory logic course of the effects on learning of an automated logic tutorial program. He received a reprimand from the Dean (Stearns) for failing to give FCEs. (The Dean later apologized.) In my years on the College and University Promotion and Tenure Committees, I saw few cases in which learning was evaluated by anything other than selected letters from students, an occasional anecdote if someone on the committee had observed a class or talked with students, and the overall instructor and course evaluations on the FCEs. The annual reports required of H&SS faculty ask about courses taught, new courses created, and overall FCE ratings, but nothing about serious evaluations of learning. Moral: *There is a Gresham's Law in teaching evaluations—FCEs drive out more serious measurements of learning.*

Educational research confirms most of these morals, and others equally dismaying. Basically similar instruments are given in hundreds, probably thousands, of colleges and universities, and they have been repeatedly studied. Studies find that average FCE scores tend to be roughly constant for an instructor across courses, that student grade expectations explain about 16% of the variance in FCE rankings—and that final grades and FCEs are correlated to about the same extent—that class size is negatively correlated with FCEs, that the "enthusiasm" and reputation of the instructor influence FCEs, and a recent study finds that the physical attractiveness of the instructors—especially male instructors—influences FCEs. (I especially object to that.) Some of these effects have been demonstrated not just by correlations but by experimental interventions of various

kinds. I know of no good studies that show that courses in which more learning actually goes on—or more that is worth learning is taught—measured, for example, by pre-test and post test performances are more highly valued by students *for that reason* than are less instructive courses. But even if that were so, the FCEs are heavily biased instruments: biased against faculty who have formal approaches, who let students know their grading will be rigorous, who aren't comely, who adopt original methods of instruction. *No student would agree to be evaluated by such criteria. No promotion committee would explicitly count such considerations against promotion of a faculty member, but implicitly it is done all the time*.

I have heard three objections to these arguments. First, that there is nothing to replace FCEs; second that without them students will have no way of communicating their collective praise or dismay with courses or instructors; and third, that CMU students are perfectly capable of accurately estimating how much they have learned in a course, and to claim otherwise insults them. The objections are without merit. Portfolios of lecture notes, syllabi, etc. and videotapes can give evidence of the quality of content and presentation; pre and post tests can give evidence of skill acquisition; student essays at the beginning and end of a course can give evidence of writing improvements; and no doubt if we troubled we could think of a variety of other ways of estimating learning, and we could begin to introduce them into assessments. Just about any CMU student in just about any afternoon can put up a web page for voluntary reports on classes and instructors. Sometimes they appear spontaneously (at one time, years ago, a student formed a site called AssassinateGlymour.alt; I lived with it.) Students can send praise or complaint to the Head or Dean, and (in my day!) before FCEs, we did. Finally, misunderstanding of what one knows, or has learned, is the human condition, and it is no insult, only truth, to say that CMU students are quite as human as everyone else.

FCEs may have some marginal value in identifying really dreadful or negligent instructors, but they exist because they are a double convenience. They allow the University to claim to students and parents and even to itself that teaching—and learning—are taken seriously, and they save the time and trouble more serious assessments would require. A Dean or Department Head or committee can glance at overall evaluations of course and instructor and form a judgment. Serious evaluation of learning is a lot more trouble, and probably a lot more intrusive. Faculty should welcome some intrusion if it is rationally aimed at assessing their effectiveness as instructors.

Susan Ambrose presented many of the objections to FCEs summarized above to the faculty and to the Faculty Senate, hoping at least to rid the new evaluation instrument of the "overall course" and "overall instructor" ratings that go into the manila folders that influence faculty careers at promotion and tenure time. False hope. The faculty, the student body and the administration alike would do better to heed her counsel and reconsider the installation of yet another Faculty Course Evaluation instrument.