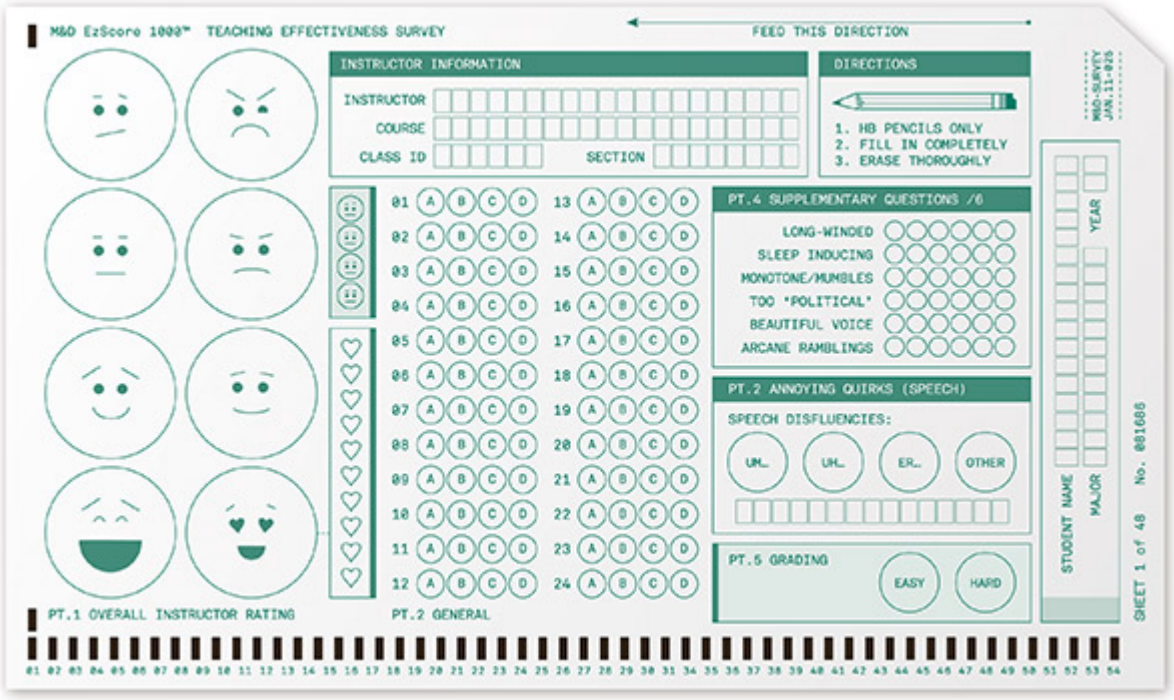
The New York Times

Report

Judgment Day



Abi Huynh/Royal Academy of Art, the Netherlands

By MARK OPPENHEIMER
Published: September 19, 2008

Annemarie Bean, who goes by Anna and is a distant, poorer cousin of the family that owns the L.L. Bean clothing business, is the kind of professor who draws students to small New England liberal-arts colleges like Wesleyan. She is funny, enthusiastic, devoted to her students and passionate about what she teaches. Her subject areas are offbeat and slightly avant-garde, the kind of stuff that students, and their ostensibly liberal faculties, are said to find thrilling: African-American theater, the history of minstrelsy, "whiteness studies" — essentially, the intersection of race and theatrical performance in modern America. Beyond her subject matter and top-notch education, including a Ph.D. from New York University's acclaimed performance-studies department, she just seems like a good fit for Wesleyan. She is an alumna of the college, class of '88; she is informal in her manner, tall and limber like a dancer, bright-eyed, the opposite of stuffy, eminently approachable; and she suggested lunch at It's Only Natural, the pride of Middletown, Conn., a regional mecca for vegetarian, vegan and macrobiotic dining. (Nothing says "Wesleyan" like lunch at It's Only Natural, where you eat bulgur wheat beneath paintings by local artists.) Bean knows that she belongs at Wesleyan, which is why she's especially sad that her students fired her.

They did not actually give her the pink slip, of course, and for that matter Bean did not receive a pink slip. A visiting professor on a one-year contract with the African-American studies department, Bean was fired by not being rehired. Before her first year of teaching, she received a letter from Renee Romano, her department chairwoman, saying that she would be recommended for a second year if she met certain benchmarks in her students' evaluations of her. Specifically, for the fall 2007 term her teaching and the overall quality of the course had to be "rated in the top two categories (Outstanding and Good) by at least 85 percent of the students in both your courses." When, at the end of the semester last December, she got only 76 percent in one of her classes and 73 percent in the other, she knew her job was in jeopardy. In January, she asked Romano if she should begin

looking for another job. She heard nothing until mid-March, when the dean, Donald Moon, still wavering, asked her to write a self-evaluation.

Finally, Bean says, Gayle Pemberton, the new chairwoman of African-American studies, told her she was out of a job — partly because, Pemberton said, Bean had not received high-enough marks in the category of "student effort," a category unmentioned in Romano's letter. According to Pemberton, not enough students had marked "strenuous" to describe their own effort in Bean's class. Put another way, Bean was being punished for her students' admitted laziness. When Bean asked Dean Moon what had happened, he referred back to the original criteria of quality of the course and quality of the teaching. Neither Moon nor Pemberton, who has since retired from Wesleyan, would speak on the record about Bean's case. A university spokesman, citing Wesleyan's policy of keeping personnel matters confidential, would say only that Bean's description of her contract "is not accurate." But Bean maintains that her students — about three-quarters of whom, after all, rated her class and teaching "good" or "outstanding" — gave the administration sufficient reason to end her time at Wesleyan.

A single mother, 42, with a 10-year-old daughter and an 8-year-old son, Bean found herself unemployed. When we met for health food in June, she was not sure what would come next. "I'm going on unemployment starting July 1," she told me. "I am selling my house in West Hartford. I have an open house tomorrow, because I can't afford the mortgage payments." In July, she and her children moved to Bennington, Vt., where they now live with her boyfriend.

On one level, Bean's case seems a simple miscarriage of justice. A highly qualified teacher and scholar was hired, received good student evaluations, but was not rehired because she failed to reach course-evaluation standards that were created seemingly at random. But it might change your opinion to know that Bean was denied tenure at her previous school, Williams College, partly because of concerns about her teaching.

And it might change matters further to know that at both schools opinion about Bean was highly polarized: many students adored her, and her classes were oversubscribed, but a small minority of students loathed her. To judge from her student evaluations, she was less an amiable mediocrity than a controversial iconoclast, striking some as a master teacher and others as an incoherent mess. "I love Professor Bean," reads one typical evaluation. "I learned very little," reads another. Or try this for a contrast: "I've probably never learned as much in any class before" versus "Bean was enthusiastic, but it was not contagious."

Whatever Anna Bean is really like in the classroom, her situation highlights the difficulties encountered every time a student is asked to evaluate a professor. Today there is hardly any college or university that does not have a formal system for soliciting student feedback about teachers. How these evaluations are used varies by school. At the top universities and elite colleges, a good research record can easily outweigh poor student evaluations in the eyes of the tenure committee. Indeed, a frequent complaint of students at the best universities is that administrators don't care whether their top faculty members can teach (or even do teach). But at most other schools, the drive to teach students well — and keep them happy and attract more applicants — has elevated the role of student opinion in the faculty's fortunes. Administrators say their forms, often filled out by students during the last class of the term, before they take a final exam or receive final grades, provide relatively objective criteria for measuring how well a school is educating its students.

Professors are more ambivalent, and they happily share theories — some supported by research — that what students are really evaluating is less pedagogy than whether a professor is funny, handsome or, above all, an easy grader. It hasn't gotten so bad that department heads are consulting popular sites like ratemyprofessors.com, where students can anonymously post any invective they want against their most despised teachers (and flag the most comely teachers with chili peppers next to their names). But Internet evaluations only wound egos; the official evaluations by students can end their teachers' careers.

According to Wilbert McKeachie, a retired psychology professor at the University of Michigan, the first teaching evaluations were most likely conceived by professors doing research on their own teaching. "I think they go back at least until the 1920s," says McKeachie, who has studied evaluations for decades. In 1945, McKeachie entered graduate school at Ann Arbor, after a wartime Navy career in which his ship was hit by a kamikaze plane off Okinawa. "My supervisor used to run a seminar for those of us who taught," McKeachie

says. “We talked about teaching, and we were beginning to use student ratings then. My dissertation compared three methods of teaching: recitation, discussion and tutorial, I called them. I used student ratings as well as measures on the final examination to assess the results.” After seeing which students fared best on their exams, then looking at what they thought of their teachers, McKeachie argued that the students who learned the most tended to give the most credit to their teachers. “It’s not just a popularity contest,” McKeachie says. “The best teachers tend to get better ratings.”

In the beginning, researchers like McKeachie were dealing with small sample sizes, using mostly questionnaires they developed themselves and administered to students at their own schools. Over time, of course, students’ evaluation of teachers became commonplace in the United States, and it has since spread abroad, slowly, to places like England. To take a random sampling, the University of California adopted a system in 1969 and Duke in the mid-1970s. Wesleyan has used student input for “more than 25 years,” according to David Pesci, the spokesman.

Are such evaluations reliable? By one estimate, there have been as many as 2,000 studies of the question. Some of these studies mocked the idea that you could even begin to assess with surveys something as subjective as good teaching. For “The Doctor Fox Lecture: A Paradigm of Educational Seduction,” a 1973 article still widely cited by critics of student evaluations, Donald Naftulin, a psychiatrist, and his co-authors asked an actor to give a lecture titled “Mathematical Game Theory as Applied to Physician Education.” The actor was a splendid speaker, his talk filled with witticisms and charming asides — but also with “irrelevant, conflicting and meaningless content.” Taking questions afterward, the silver-haired actor playing “Dr. Myron L. Fox” affably answered questions using “double talk, non sequiturs, neologisms and contradictory statements.” The talk was given three times: twice to audiences of psychiatrists, psychologists and social workers, the last time to graduate students in educational philosophy. In each case, the evaluations by the audience were highly laudatory. To these audiences, Dr. Fox was apparently articulate and intellectual, not a fraud.

Of course, we all enjoy spending time in the company of charismatic people like Dr. Fox, and we’re more sympathetic to what they have to say (indeed, that enhanced attention people pay comes close to a plausible definition of charisma). So the Naftulin study, although more in the realm of useful anecdote than of statistical analysis, at least shows that questionnaires might measure something related to good teaching. You may be far more pessimistic about evaluations after reading “Fudging the Numbers,” an article published last year in Teaching of Psychology. For this study, Robert J. Youmans and Benjamin D. Jee, who were then graduate students at the University of Illinois at Chicago, asked six discussion sections of a course to evaluate the professor — but students in three sections were offered chocolate before they completed the forms, and those students gave more positive evaluations. It was a small study, but it was enough for the authors to sound a note of caution about the influence of extraneous factors on students writing evaluations. You wonder: If it’s raining outside during the last class, will students feel more negative as they jot down their impressions of the course? What if the football team just lost a big game?

Critics of student evaluations have many complaints: lazy department heads use them instead of visiting classrooms and judging for themselves who is teaching well; students have a bias against hard classes (or math classes, or early-morning classes); having students fill out evaluation forms, often at midterm and at the end of the term, wastes class time. But the great concern of those who evaluate evaluations is that professors buy favor not with their charm, nor with their chocolate, but with a far more valuable medium of exchange: grades.

In his polemic “Why the University Should Abolish Faculty Course Evaluations,” published in a faculty newsletter in 2004, Clark Glymour, a philosopher at Carnegie Mellon University, argues that giving good grades can even make up for a professor’s lack of charm and wit. When Glymour was department chairman in the 1980s, a “newly hired assistant professor consistently received the lowest faculty course evaluations in the department, and I was concerned for his career,” Glymour writes. “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.” Traditional and serious, lacking the levity students appreciate, the professor refused to seek help with his teaching, but he assured Glymour that his student evaluations would nevertheless improve significantly. “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.' "

In faculty lounges, anecdotes like these are as common as Obama supporters. But among administrators, the conventional wisdom is quite different. "All the evidence," says Steven Olswang, a professor of education and the interim chancellor of the University of Washington, "shows that student evaluations correlate with all other measures: peer evaluations, outcome evaluations. If they're good, their student evaluations are probably good." Evaluation enthusiasts point to studies by researchers like John Centra, who designed an evaluation form sold by the Educational Testing Service to hundreds of schools. For one study, Centra looked at evaluations from more than 5,000 classes, trying to discern if students who rated their teachers highly at term's end did so because they were expecting high grades or because they believed they learned a lot. "What we hoped to find," he says, "and we did, is that where there was good learning, there were good ratings, and expected grades were higher as well." Controlling for the expected grades, Centra concluded it was students' beliefs that they'd learned a lot that prompted good evaluations.

But these studies have dangerous pitfalls. After all, when students report having learned a lot, how do we know to trust them? A certain kind of teacher might give students the impression they have been well educated, to the point where even if she gives them bad grades, they still praise her teaching.

To know what's an effect of grades and what's an effect of learning, you need to be able to measure them separately. That's what Bruce Weinberg, an economist at Ohio State University, and two colleagues sought to do in their innovative 2007 study. Weinberg asked the Ohio State registrar for a decade's worth of students' grades and teacher evaluations, from 1995 to 2004, in both introductory and higher-level economics classes. Because he could assess how well students actually did in future classes, Weinberg had a better assessment of how much students learned than just what they wrote on their evaluations.

Weinberg's findings gave pause. It was true that students who gave positive evaluations to their introductory-class professor were more likely to do well in future classes. But the reassuring news stopped there. When Weinberg crunched the numbers further (controlling for grades), he found that those who did well in the higher-level classes were not more likely to have given their introductory-class professor a positive evaluation. In other words, when you use performance in higher classes as the measurement of learning in the previous class, the correlation between learning and positive evaluations breaks down; it becomes plausible that it was the grade, not the actual learning, that got the kids so psyched about their freshman econ professor.

Even if the optimists were right, and teaching evaluations could help schools divine which professors were best at communicating the facts of freshman biology, there remains a big problem: the sciences are not the humanities. We know how much basic chemistry students need to know before enrolling in organic chemistry, but it is far more debatable which facts, skills and habits of mind a teacher of black theatrical history ought to convey.

This conundrum surely accounts for some of the murkiness surrounding the case of Anna Bean. She says she believes that part of her job is to discomfit students, to rid them of easy assumptions (for example, that being white, as she is, is the norm while everyone else is a minority). And in principle most professors would agree this is a laudable goal. But students don't always want to buy what teachers think they're selling. In their 2006 article, "My Professor Is a Partisan Hack," the political scientists Matthew Woessner and April Kelly-Woessner, who obtained course evaluations from almost 1,400 students at 29 colleges, found that political-science students give poorer evaluations to professors whose perceived political views they disagree with. "Students even report they learn less from professors whose views are different from their own," Kelly-Woessner says. "That's counterintuitive. You'd expect that students would learn more from people with different ideas. But what the political psychologists say is that people tune out those who make them uncomfortable. It's like why liberals don't listen to Rush Limbaugh. Students believe they learn more from people who say what they say." Kelly-Woessner found that the bias works against liberal and conservative professors almost equally. "There's some expectation today that a professor be objective or evenhanded, and if professors violate those norms, they can pay a price for it in student evaluations," she says.

Professors across the political spectrum may pay that price. But Carolyn Byerly, a former journalism professor at Ithaca College, argues that her radical views and focus on how race, gender and sexual orientation are handled by the media led students to destroy her chance for tenure. “I had submitted my tenure file — and I also included my evaluations, which included nine consistently excellent peer reviews from colleagues who observed my teaching,” Byerly says. “Then came the teaching evaluations. Overall they were excellent from my students, but those anonymous, handwritten evaluations were singled out by my dean and chair, and 43 of several hundred were selected as indications that I had not met the standard of excellence for teaching. Almost every one of the 43 were full of gender bias: ‘This teacher has a political agenda.’ ‘This teacher supports gay rights.’ ”

In one of the few lawsuits ever brought over student evaluations, Byerly sued Ithaca in 2001 for sex discrimination. She lost on summary judgment, and her subsequent appeal was denied by “three white male judges, one of them 87 years old, who didn’t understand why things like feminism and race and gender issues had anything to do with why I was brought there to teach journalism,” she says. (Byerly is white herself.) Ithaca, for its part, would only say through a spokesman that her claims of sex bias “did not have merit.” And to be fair, it’s hardly clear that Byerly’s sex was at the root of the school’s problem with her. She seems to admit as much when she describes the culture of the school. “Ithaca, campuswide, is almost exclusively white, almost exclusively upper middle class,” Byerly says (exaggerating slightly on both counts). “Somebody like me, who comes in and says we have to question things like class privilege and look at the ways the news is covering gender and race — someone like me is going to have a more difficult time getting along.”

Academic administrators want many things, from good pedagogy to clean campuses to successful athletic teams. Among the things they want most is for everybody to get along. One obvious way they learn about dissension is through student evaluations, especially in classes where the subject matter might allow professors to air personal, possibly radical opinions. Byerly’s evaluations at Ithaca showed that some of her students, an opinionated minority, felt no more kindly toward their white professor than she did toward the white, rich lot of them. She has a complicated explanation for why her department chairman and dean didn’t like her, which involves the “hegemony” of certain corporate interests over the school. Whether or not that’s entirely true, she could still be right that the school wasn’t interested in her brand of liberalism. And instead of saying as much, Byerly’s bosses had students, through their evaluations, do the talking.

I have taught college students, and read evaluations of me, but I was unprepared for the naughty thrill of reading evaluations of somebody else. Getting to read what Anna Bean’s students thought of her was like finding your neighbor’s bank statements, or maybe medical files, on the sidewalk.

Bean prepared me for what to expect. At both schools where she taught, many students adored her. (When she was denied tenure at Williams, many alumni wrote letters in protest.) Others, of course, were indifferent or lukewarm. And a small minority couldn’t stand her. When I looked through evaluations of Bean’s class, Blackface Minstrelsy, Then and Now, I found all three groups. The great: “I found the teaching amazing. . . . I believe that the love and expertise that the professor obviously has in the subjects shines through in her teaching.” The good: “Teacher is well informed and has interesting topics.” And the very, very bad: “In general this course deteriorated and by the end of the class we weren’t even talking about minstrelsy. Tremendous amounts of time were wasted by Bean’s lateness (due to yoga class), absence or even inability to operate technology. . . . I found her completely unstimulating and unable to lead productive classes.”

If you came across the whole pile of evaluations on the sidewalk, you’d form a picture of a somewhat disorganized, technologically inept, very learned, passionate teacher — an acquired taste. It would be clear that her particular cocktail of traits was very appealing to some students, the ones who loved her passion or her subject matter so much that they didn’t think her tendency to be late or frazzled was worth mentioning. You’d see that other students, meanwhile, were unmoved by her considerable energy and deep knowledge — instead, they felt abused by her politics, her scattered style or her deviations from the syllabus.

Bean told me that she had a good sense of who had written the most negative evaluations. “I found there was a small group of mostly white men,” she said, “who sat there the whole time wearing their white hats on backward, sitting there angrily, who didn’t like the class.” The stereotype Bean was invoking is well known to recent college alumni, especially of wealthy Northeastern schools. There is a look popular among athletes and

their hangers-on, who wear white baseball caps with the name of a college embroidered above the brim. When you see those boys in class, you do figure — at least I always do — that if they're not jocks, they're part of a jockish, frat-boy scene. On a campus like Wesleyan, these are the boys who have not bought into its famously liberal culture. And if you're Anna Bean, and you're teaching classes called Whiteness or Blackface Minstrelsy, you worry, despite your best efforts, that they might be suspicious of what you have to say.

Where Anna Bean saw "white men . . . wearing their white hats," Caroline Byerly saw an "almost exclusively white, almost exclusively upper-middle- class" campus. In each case, their caricatures of the students they were paid to inspire probably say something truthful, if crudely expressed. There was a certain kind of student, they knew, who simply was not going to like Carolyn Byerly, the outspoken lefty, or Anna Bean, the WASP aficionado of blackface history. And there's a certain kind of class material that is bound to elicit mixed reviews, especially if a teacher lacks a deft classroom touch. When one student wrote that "Bean constantly referenced certain individuals in the class in inappropriate ways," I couldn't decide who those students might have been. Were they minorities whom Bean asked about their own life experiences? Were they white-hatted athletes whom Bean singled out to provoke? Whichever it was, at least one student felt that "the class dynamic was not only uncomfortable but never addressed."

Reading some students' overblown praise, and others' righteous anger, made me crave objective criteria for evaluating teachers. But for Bean's and Byerly's classes, there is no way to know what criteria to use. It would be impossible to perform Weinberg's study on the classes that Byerly and Bean teach. We could never agree on which higher-level classes would measure how much students learned the previous semester — in what class can you demonstrate the skills mastered in Blackface Minstrelsy?

If there's no consensus about how well evaluations work in a class like basic microeconomics, it's even more difficult to know how seriously to take them in classes like Byerly's or Bean's. An administrator must synthesize multiple ways of looking at a humanities professor — one who was given no set syllabus and no canon of knowledge to convey, just a simple charge to develop students' minds in ways they might only appreciate decades later but are asked to describe in 20 minutes carved from the last class before vacation. If the evaluations themselves are subjective, so, too, is reading them; no matter what you think of Anna Bean, it's hard to be unmoved by one student's poignant critique that she "made 'jokes' about how these evaluations will influence her position here and her children's health care."

It's unlikely that another 2,000 studies will end the debate over student evaluations. Just as students like the professors who grade them well, professors like the evaluations on which they do well. You might think, for example, that Carolyn Byerly would be an unwavering opponent of student evaluations, but it turns out that she's an optimist about their potential for good. Her epiphany came after she was hired by Howard University, the historically black school in Washington. There, her student evaluations have been positive from the start. "At Howard," she says, "I am above the national mean in almost every single category of evaluation — the clarity of my material, organization, ability to communicate information, a bunch of things." Howard uses the latest version of John Centra's fill-in-the-bubble test sold by the Educational Testing Service, the corporate behemoth of the testing field. Based partly on her popularity as measured by that evaluation form, Byerly was granted tenure last year. "I personally believe that students ought to be able to evaluate their faculty members," she says.

Having died by the system, Byerly has now been granted new life by it. But to many, it's the system itself that is choking higher education. When students in the 1960s demanded more say in academic governance, they could not have predicted that their children would play so outsize a role in deciding which professors were fit to teach them. Once there was a student revolution, which then begat a consumer revolution, and along with more variety in the food court and dorm rooms wired for cable, it brought the curious phenomenon of students grading their graders. Whether students are learning more, it's hard to say. But whatever they believe, they're asked to say it.

Mark Oppenheimer is director of the Yale Journalism Initiative and an editor of New Haven Review. He last wrote for the magazine about the New Age publisher Louise Hay.

Source: http://www.nytimes.com/2008/09/21/magazine/21wwln-evaluations-t.html?ref=magazine