

## STUDENTS MUST EVALUATE FACULTY

David P. Doane

### *Students Matter*

Of course students should evaluate faculty. Who else is better qualified? Self-respecting persons or enterprises evaluate themselves both through introspection and by feedback from external observers. Evaluation is attempted despite awareness that any evaluation is imperfect. Why students? Consider a self-test question from a chapter on quality improvement in a recent textbook (Doane, Mathieson, Tracy, 2001).

- Quality in a product is *best assessed* by
- a. trained statisticians.
  - b. skilled mechanical engineers.
  - c. attorneys who handle product liability.
  - d. government quality inspectors.
  - e. customers.

It is a tenet of quality improvement that the consumer is the only observer who sees the whole elephant. The agents named in (a) through (d) also assess quality. But it is the customer who pays. Granted, employers, graduate schools, and other faculty are also consumers of what goes on in a classroom. If a university has the time, resources, and ingenuity to solicit their views, fine. But students write a check. They pay us and are old enough to choose a President. It would be arrogant to deny the validity of their views. The challenge to the academy is to design credible evaluation methods that pursue worthwhile ends, and to be wise enough to use student input sensibly.

### *Why Not Faculty?*

There may be a role for faculty peer evaluations. But faculty lack the time to visit classes on a daily basis. Will they obtain a representative sample? Are peers more objective than students? Would turf issues, personality conflicts, or differences in teaching philosophy (Olsen-Fazi, 2004) color peer evaluations? Are not faculty brethren already “converted” to a certain level and style of discourse? I recall a student’s comment when looking at the equation for the expected value of  $Y$  for a given value of  $X$  (i.e., the conditional mean of  $Y$ ):

$$E(Y|X) = \beta_0 + \beta_1 X.$$

The student asked, “Why is  $Y$  expected, and why is  $X$  given?” A colleague might laugh at that question. But to some learners, an equation is neither an explanation nor a statement, but may actually be a barrier to understanding. A resourceful teacher might craft alternative illustrations to get the idea across, using diagrams, scatter plots, or spreadsheet simulations. Would a highly-trained peer who understands mathematics realize that a creative alternative was overlooked if the instructor merely referred the student to page 622 of the textbook?

For that matter, are faculty ever likely to disparage a peer’s teaching? My 35 years’ experience with peer evaluations (including service as department chair as well as time spent on numerous faculty review committees) suggest that peer reviews tend strongly toward encomia sprinkled with just enough mild self-help suggestions to establish a flavor of objectivity.

If student evaluations are sought, the survey instrument ought to be simple, with both scaled questions and open-ended responses. Specialists who support lengthy instruments scales have some arguments on their side, but lower response rates and less thoughtful student comments may be the price. If we are not trying to explain or even necessarily measure a single overall construct called “good teaching,” we can focus on a few relevant, modifiable behaviors that a student can observe reliably. If there is no “bottom line” question (e.g., “Would you recommend this instructor?”) administrators are very likely to invent one by summing the scales. This being the case, I feel that such a question might as well be included, on the grounds that it is better than the alternative.

### ***To What End?***

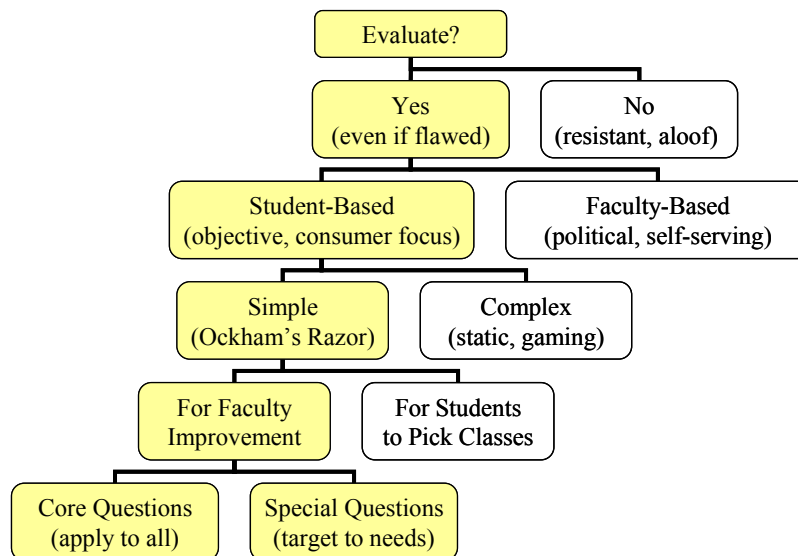
Students already have devised ways of posting information on the web. One cannot feel confident that this information is reliable, or that it will always be used responsibly. Yet even critics of student teaching evaluations (e.g., Gray and Bergmann, 2003, p. 45) concede that students can recognize extremes of teaching quality. As the *Oakland Undiapered* experience proved, students *will* evaluate us. Once we accept this, we can turn to creating a valid evaluation instrument that addresses our own goals (*not* just to help students pick an instructor) and to using it in sensible ways.

The never-ending quest for faculty self-improvement should be the main goal of student evaluations (not to provide students with information to pick instructors). If some information is shared with students, we can at least attest to its accuracy and completeness. If some information is used in the faculty promotion and tenure process, we can design safeguards to prevent undue reliance on it. Many academic units already strive to assemble diversified teaching assessment portfolios (e.g., self-statements, syllabi, assignments, exams, samples of student work, alumni letters) as recommended by experts (Centra, 1979; Seldin

and Associates, 1990). Misuse of student evaluations is a criticism of the academy, not of the evaluations, just as misuse of taxes is a criticism of government, not an argument for abolishing taxation.

### ***Design and Administration***

Figure 1 summarizes the argument. If we seek mainly to provide relevant feedback to help faculty become more effective teachers, we should seek (1) core questions that were sufficiently cross-cultural to apply across all disciplines, class levels, and instructor types, and (2) special questions specific to departments, disciplines, class levels, and even instructor types. Labs, lectures, case projects, and musical or artistic performances may require unique evaluation methods. We might also want to give targeted feedback to new, inexperienced faculty or part-time, adjunct, and distance-learning faculty. With web-based surveys, questions can even be tailored to individuals.



**Figure 1 Flow Chart of Reasoning**

### ***To Web or Not to Web?***

The time has probably come to use the economies of scale of the web to administer evaluations under standard, uniform conditions. Many universities lack the time, money, or will to distribute paper surveys to all classes at the same point of time. Despite the criticisms of web-based evaluations (e.g., possible response biases) they may be no worse than present methods. The web may help avoid

instructor gaming (e.g., choosing the time of administration to coincide with good news or low attendance) by being open for an entire week (or even a midterm evaluation). Web surveys may be better for non-standard students (e.g., distance learners) or those who choose to respond at odd times. The web also permits immediate feedback to the instructor. However, issues of response rates and privacy will require study, and it is well not to be sanguine about them.

### ***A Bit of History***

Academics are familiar with the requirement that any evaluation instrument or system must possess both *reliability* (consistently giving results) and *validity* (measuring what it's supposed to measure). Academic empires and institutes have arisen dedicated to perfecting the mechanics of measurement and validation. One further requirement that is sometimes forgotten is *face validity* (satisfying people's beliefs about fairness).

In the early 1970's, some faculty aligned themselves with students who were interested in establishing a reliable system for evaluating classes and teachers. The *Oakland Undiapered* was a hit-or-miss enterprise that suffered from lack of continuity in its administration, and which received only indirect institutional support in the form of access to facilities and subsidized printing. It was a student-run operation that depended on the dedication and goodwill of students who took an active interest in the process. But students faced conflicts of time that caused imperfect class coverage (some said highly imperfect) and subjective editing of student comments. These student Pioneers (who were not yet Grizzlies) eventually graduated and turned the enterprise over to more or less willing successors. The *Undiapered* did not last long, though it left its mark.

Faculty members Robbin Hough, Don Hildum, and several others became excited about the possibility of developing a novel instrument to assess teaching. Many academics would have emulated best-practice methods that evolved at large universities, embodying thousands of person-hours by specialists in design, scaling, and educational measurement. Being a truly original thinker, Hough preferred to start from scratch, and to try to find out what *Oakland students* thought was important about good teaching. His approach was direct. He and colleagues across the university, with the help of student volunteers, undertook a university-wide sample of hundreds of students in a cross-section of classes across all levels. Each student in the participating classes was given each a blank sheet of paper containing one question:

“Please list attributes of a good teacher. List as many as you wish.”

This approach was consistent with the Oakland that existed in those days, where students and faculty shared the perception that we were innovators who were, by golly, going to do everything “our way.” Since Oakland’s genesis was a reaction against the highly-codified learning environments of large public universities, it was easy to believe that our students really might see their relationship with teachers in a different way.

Using scissors, those comments were cut out on slips of paper (or copied to small slips of paper if necessary) and placed on a big table in North Foundation Hall where the SBA housed its computer lab. Faculty volunteers then sat down with these (thousands?) of slips of paper and attempted to group the comments in piles representing what we felt was the same attribute. It was a kind of subjective factor analysis. When piles began to seem heterogeneous, they were split. When piles seemed similar, they were combined. We went through each pile again, moving attributes to other piles and sometimes creating new piles. This process took days. Eventually, we ended up with perhaps 30 piles. Each pile was assigned a one-sentence description that the faculty felt captured the attribute.

Next, another university-wide cross-sectional sample was surveyed, in which respondents were given a sheet of paper listing the attributes in random order. The student respondents were asked to circle the five most important attributes and to write an *X* beside the five least important. The attributes were ranked according to circles and *X*’s results were tabulated, and a weighting process was used to select the attributes that were most often circled and least often circled. About 15 items were selected for further testing. Using a standard 5-point scale, we began using this instrument to evaluate teachers. Regression analysis was used to confirm that the most-circled survey items were in fact predictive of the “bottom line” instructor rating. None of us thought to try to publish the results. Perhaps no journal would have wanted it, but we probably should have tried.

This research formed the basis for the student evaluation system used by the SBA for the next 20 years or so. Variants of this system were used by SECS and perhaps by other academic units. The SBA survey has gradually been modified through *ad hoc* processes that bear little resemblance to the one described above. SBA most recently trimmed the length of its survey, partly because factor analysis showed overlap in many of the scale items. Now SBA is experimenting with a web-based system, as are several other schools (SECS, SHES). It is too early to tell whether these web-based systems will be “keepers.” But the underlying issues are no different than in the past.

### ***What Are the Alternatives?***

I leave it to others to make the case for alternatives to student evaluations. Although scales now in use have not been thoroughly studied, there is no doubt

that student evaluations can be statistically valid (e.g., McKeachie, 2004, Seldin, 1980). While I agree with critics that “a few decimal points difference” do not indicate any real difference in teaching effectiveness, my experience as a department chair convinces me that broad groupings such as quintiles are useful in showing individuals how they are perceived by students, and in helping departments assign faculty to classes where they are most effective. I think it is inevitable that departments will seek to reward instructors who are perceived favorably by students (and conversely) and to consider an instructor’s willingness and ability to address teaching issues in tenure and promotion decisions. The real issues are (1) creating valid, reliable, and credible student evaluations; (2) the weights we attach to student evaluations; (3) broadening the assessment of teaching; and (4) the role of teaching evaluations in administrative matters.

### ***References***

- Centra, John A., *Determining Faculty Effectiveness* (1980), Jossey-Bass
- Doane, David P., Kieran Mathieson, and Ronald L. Tracy (2001), *Visual Statistics 2/e*, (Irwin/McGraw-Hill), p. 425.
- Gray, Mary, and Barbara R. Bergmann (2003), *Academe*, Vol. 89, No. 5, p. 45.
- McKeachie, W. J. (2004), *Academe*, Vol. 90, No. 1, p. 6.
- Olsen-Fazi, Annette (2004), *Academe*, Vol. 90, No. 1, p. 4.
- Seldin, Peter (1980), *Successful Faculty Evaluation Programs* (Coventry Press), pp. 36-65
- Seldin, Peter and Associates (1990), *How Administrators Can Improve Teaching* (Jossey-Bass), pp. 89-103