0605-1

Making Better Use of Student Evaluations of Teachers

Peter Elbow

There is a widespread skepticism about students as evaluators of teachers. I hear these criticisms most frequently:

- Students are immature and not yet educated and don't know about teaching and learning as we do.
- Students just go on feelings, what they like, what's fun or entertaining; they can be seduced by a good show and easy grades.
- Student estimations of teachers vary wildly; their evaluations obviously have no reliability.

What shall we conclude from these charges? What I conclude is this: we feel more keenly the problems of evaluation when we are on the receiving end than when we're on the giving end. That is, the charges I've just summarized simply throw a clearer light on the problems in *all* evaluation, particularly conventional faculty grading of students. Let me set the problems of student evaluations into a larger context by briefly standing back and talking about evaluation in general.

Trustworthy, fair evaluation means giving God's verdict—finding the single verdict that all right-minded, good readers would agree on. The problem is that God isn't telling her verdict, and we cannot get readers to agree—not even *good* readers. It may sound extreme to invoke God here, but we can't be cavalier about evaluation in education. A single student's evaluation of a teacher doesn't carry much weight, but a single teacher's grade for a student often carries a lot—for example, in an application for a scholarship or a job or professional school. We can't just take a fashionably theoretical view of the grades we give: "Oh well, of course my grades are 'situated' and 'interested'—so what else is new?" Because grades carry heavy consequences, we cannot take anything less than genuine fairness as our goal—God's view, correctness—yet we know that trustworthy, fair evaluation is not possible.

I'm not saying anything new. We've long seen this predicament on many fronts. Research in evaluation has repeatedly shown that if we give a paper to a set of readers, they tend to give it the full range of grades (see Diederich for a classic exploration; for an indication of how long people have noticed this problem, see Starch and Elliott [three citations] and the summary of their work in Kirschenbaum, Sidney, and Napier 258–59).

We know the same thing from literary criticism and theory. The best critics disagree about the quality of texts—even about what texts mean—and nothing in literary or philosophical theory gives us any agreed-on rules for settling such disputes. Barbara Herrnstein Smith may not be too cynical in concluding that whenever we have interreader reliability, we have something fishy. And students know the same thing from their controlled experiments of handing the same paper to different teachers and getting different grades. (Perhaps this explains why we tend to hate it when students ask their favorite question, "What do you want for an A?": it rubs our noses in the unreliability of our grades.)

Champions of holistic scoring will reply that they do get readers to agree, but they get that agreement by "training" the readers before and during the scoring sessions—that is, by getting them to stop using the conflicting criteria and standards they normally use outside the scoring sessions. Thus the reliability in holistic scoring measures not how texts are valued in natural settings by actual readers but only how they are valued in artificial settings with imposed agreements.

Nevertheless, these practical and theoretical problems don't permit us to decide we can get along without evaluation. Everyone seems to agree that we need some kind of evaluation of students by teachers. And even cynical teachers realize that we can't just refuse to have our students evaluate us. Those who ignore the other reasons still must acknowledge that colleges operate in a

The author is Professor of English at the University of Massachusetts, Amherst. This paper was presented at the 1991 ADE Eastern Summer Seminar at Skidmore College. A version of this paper will appear in a volume on evaluating teachers of writing, to be published by NCTE.

competitive marketplace where students are consumers—and are scarce and precious. If we are selling a service, we can't say, "Let's ignore the consumer's opinion." We need students more than they need us; it's a buyer's market; they can always go elsewhere. And this asymmetry is also structural: students can learn without teachers, but teachers cannot teach without students.

So if fair, trustworthy evaluation is impossible but evaluation is necessary, what is the moral? I see only one answer. We should do less of it—so that when we really *must* do it, we can do it better.

Doing it better. What would better evaluation look like? Above all, it would be more trustworthy and more informative. These two goals point straight at the main culprit in evaluation: fine-grain, holistic, numerical ranking along a single dimension. It is this alleged measurement of a complex performance—scoring from 1 to 5 or F to A—that is least trustworthy and least informative. Holistic scores are nothing but points along a yea-boo continuum—with no information about the criteria behind the yeas and boos. They tell us with laughable precision how well or poorly evaluators thought someone did but nothing about what the scorers meant by "well" or "poorly"—what they were looking at or looking for.

When we stop pretending to measure a complex performance along one numerical scale, we naturally bring in more useful evaluative information. That is, we are led to create student evaluations that focus on questions such as these: how well does this teacher conduct discussions, give lectures, devise assignments, comment on papers, help you understand the course concepts and information, help you learn to think for yourself, establish good relations with you, and so forth. Even if we are forced to ask for *numerical* answers to these questions because there are too many students to permit us to read written answers, we don't pretend we can add up these numbers and come out with a "score" for how good the teacher is. We realize that the results need interpretation.

What I am talking about here is the crucial distinction between *ranking* and *judging*. To rank is to give a single, holistic, quantitative verdict along one dimension—even though teaching is a complex, multidimensional performance. To judge is to look carefully enough at the performance to distinguish among parts or features or dimensions and decide which parts of the teaching are more effective and which less. The process itself of judging, because it is discriminating or analytic, helps us acknowledge that different dimensions of the teaching will matter more to different students. For example, some students will see a good teacher as someone who gets the material across clearly and doesn't disturb their routines or assumptions. Other students will see a good teacher as someone who shakes things up and causes them to question their routines and assumptions.

So where ranking gives us nothing but a number, judging gives useful information about which features of the teaching worked better or worse for which students. The judging process also nudges the student evaluators themselves toward being more thoughtful and discriminating about the different dimensions of teaching and learning. In contrast, ranking simply invites students to record an overall feeling with a single number. Students—indeed all evaluators—need to be encouraged to step outside merely global feelings of approval or disapproval. The most useful and interesting question in evaluation is always, What do you see?—not, How do you rank it? As C. S. Lewis puts it, "People are obviously far more anxious to express their approval and disapproval of things than to describe them" (7).

IQ scores give a vivid illustration of the ranking problem. It is plain that IQ scoring does not represent a commitment to looking carefully at people's intelligence—for when we do that, we see different and frequently uncorrelated *kinds* or *dimensions* of intelligence (Gardner). IQ scoring represents rather our culture's hunger to rank people along a single scale, a hunger for pecking orders, or, in the military metaphor, for knowing who you can kick and who you have to salute. ("Ten," mutter the chaps when seeing a beautiful woman.) We see the same principle at work in conventional grading: the use of single numbers on a one-dimensional scale to describe a multidimensional performance—with no stated criteria or categories. My argument is not against evaluation itself, only against that crude, oversimple way of *representing* evaluation—distorting it, really—as a single unreliable number.[1]

Yet I am not saying we can get rid of *all* bottom-line, holistic verdicts. After all, for the sake of important decisions in areas such as hiring, promotion, tenure, and merit pay, we often need to make the best estimate we can about who is an excellent teacher and who is a genuinely poor or irresponsible one. Such decisions can never be wholly trustworthy, but they are not so problematic as fine-grain rankings in the middle range. That is, when we look at answers to the more substantive and analytic student evaluations I've just described, most teachers will fall somewhere in the middle and get a mixed bag of results: a combination of strong-, middle-, and weak-rated features and probably many divided opinions or disagreements among students. But a *few* teachers will get strikingly strong or strikingly weak responses from many or most students in many or most categories. When we get unusually near unanimity this way—and it is supported by other evidence—we are as justified as we can be in reaching a bottom-line, holistic verdict that a teacher is excellent or poor. But we have no such justification for fine-grain, holistic, numerical verdicts in the middle.

My point is that we can never have genuine reliability—genuinely trustworthy ranking. But we can get rid of as much untrustworthiness as possible. And since we don't need to give a prize or to deny promotion to any of the faculty members in the middle, we don't need to have bottom-line scores for them. What we get instead is a lot of responses to look at and an occasion for the faculty member to talk with the chair or a committee about what that teacher does well and not so well—and about how he or she might teach better.

We can get along with much less official, high-stakes evaluation of teachers by students if we make more use of low-stakes evaluation—unofficial feedback for the teachers' eyes alone.

Doing it less. Good evaluation is more work, but less evaluation would in fact be a blessing. In particular, we can get along with much less official, careful, high-stakes, institutional evaluation of teachers by students if we make more use of low-stakes evaluation—of informal, unofficial feedback for the teachers' eyes alone. (We can also use comparable private, low-stakes evaluation from a friendly colleague or from someone in a faculty development office.) Teachers tend to learn and improve more with this kind of unofficial feedback because it is less threatening: they have more control over it and don't have to defend against it as much as they do against official, institutional feedback. (I say this on the basis of having set up a faculty peer-feedback system and served as a "visitor" in it; see Elbow.)

When I ask students for this informal evaluation, I like to do so at midsemester and to make sure they can feel that the request comes from me as teacher, not from some larger, impersonal institutional enterprise. Indeed, I often simply ask my students to write me a letter that answers questions like these: What are the most important skills and contents you have learned? What skills or abilities do you see me most trying to teach? Which features of the course and my teaching have worked well and which ones not so well? There are many benefits from this midsemester procedure: there's still time for improvement in that very semester. And the mere fact that I make the request improves my relationship with students since they see me as asking for feedback; the exercise encourages honesty and attention to teaching by them and me.

Even though this informal evaluation is private and noninstitutional, the department or chair or institution can, indeed should, make it happen—for example, by requiring faculty members to write reflective self-assessments of their teaching in which they discuss what they learned from private evaluations by colleagues and students.

Someone might object that in a section titled "doing it less" I am calling for a lot of evaluation. True enough, but this informal evaluation is easy and nonbureaucratic, and it permits much *less* official, institutional, "judging" evaluation, which could perhaps be done every two to three semesters for untenured faculty members and every four to five semesters for tenured.

Doing it. So how would they work, these official student evaluations of teachers? I would like to suggest some procedures by way of trying to answer the objections or misgivings about student evaluations that I mentioned at the outset.

- Students are immature and not yet educated and don't know about teaching and learning as we do. It's true that students may not understand the thoughts—even the goals and intentions—of a faculty member. But students know more than anyone else about the results of those intentions and goals. And it's true that students may be mistaken about their learning—they may even lie. But they see more of the teacher than any visitor possibly could. They have more evidence, more data, for they see lots of *other* teachers in just as much detail, so they are in an ideal position to make informed comparisons about the effectiveness of different procedures. Students know more than most of us about the success of different styles of and approaches to teaching, since we usually see only our own teaching and a tiny bit of others'. (An important general problem in much evaluation is the "COIK" factor: *c*lear *o*nly *if k*nown. Many explanations, lectures, classes, essays, and books seem admirably clear to colleague evaluators, but only because these professionals already understand what is being explained. When the performance is evaluated by the *intended audience,* who do not already know the material, a very different verdict often emerges. Notice, for example, that a hierarchical, deductive, abstract presentation is often just right for summing up a body of material that you already understand—but just wrong for introducing material that is new and difficult for you to understand.)

In short, the problem with students as evaluators is not whether they have useful information or knowledge, but how to get their information and knowledge in a trustworthy form. I turn now to this question.

• The most serious charge is that students just go on feelings and the pleasure principle, on what's fun or entertaining; they can be seduced by a good show and easy grades. There's a blanket answer to this criticism—namely, that pleasure and feelings are not so bad. In the academic world we suffer from a prejudice against what is easy, popular, and (the worst of all student words) "fun." There is no necessary conflict between something being easy and fun and also producing good learning. Most people learn better when they enjoy themselves.

I think my blanket answer is important, but I acknowledge that of course pleasure is not enough. We all know some teachers who are easy and entertaining but don't teach so well and some who are forbidding and no fun but in fact teach very well—and surely students must be tempted to rank the former higher than we wish and the latter lower than we wish. (But Robert Boice, in a short, useful piece reprinted as an appendix to this article, suggests the contrary, summarizing research that finds that "heavy work loads correlate *positively* with SETs," that is, with favorable student evaluations of teachers.)

A tilt toward easiness can never be completely removed, but it can be substantially diminished, to the point where we can put considerable trust in student evaluations of teachers. For the main problem here is *ranking.* We are dumb designers of evaluation unless we ask students lots of questions that have little ranking or evaluative dimension. That is, student judgments (like all judgments) are most valuable when they contain lots of *description* and least valuable when they give nothing but a number to express a degree of approval or disapproval. We can ask questions like these: How many tests were there, and what do you see them most trying to teach? How many papers were there, and what do you see them most getting at? What are the most important things you feel you've learned? What do you see as the most helpful or interesting class or activity? What do you see as the teacher's major and subsidiary goals or priorities? Do you see more emphasis on information, concepts, skills, or attitudes? How difficult was the work? What side of you did the course and teacher tend to bring out?

Then of course we can provide questions that invite rankings—but rankings only about specific criteria or particular dimensions of the course and the teacher's performance. For example, How would you rate your teacher on knowledge of the subject? guiding discussions? lecturing? devising paper topics? commenting on papers? choosing readings? relating with students?

I'm not saying that we should remove all opportunity for holistic verdicts by students about their teachers. Indeed, there is some evidence that if we give students an opportunity to say how good or bad they think a teacher was overall—if we even ask them how much they *liked* or *disliked* a teacher or a course—their holistic feelings are somewhat less likely to color their answers to more particular questions on substantive criteria. For example, global, yea-boo questions increase the chances that a student who hates a teacher can still go on to acknowledge merit in, say, the teacher's paper assignments or lectures. (Thus these questions should perhaps come near the beginning of the form.)

• Student estimations of teachers vary wildly; there is obviously no reliability in student evaluations. My response to this charge is to turn the tables and say, "Of course." Surely one of the reasons why we so often distrust student evaluations of us is that the disagreement in them—their "lack of reliability"—calls too uncomfortably to mind an obvious fact that we hide in our own grading—namely, that the disagreement we accept in published literary criticism means that multiple teachers grading the same student performance would disagree as much as do multiple students evaluating the same teacher. The unreliability of teacher grades is effectively disguised by our handy custom of only getting one opinion.[2]

Some logistics. It is important that students *write out* some of their answers instead of just checking boxes on a computer form. Writing takes a bit more time, but that's good: it adds seriousness and makes students more thoughtful than they are when they just check boxes. The only reason to have a quantified computer system is to reproduce the single-dimensional ranking that is obviously flawed (as IQ scoring and conventional grading are). There is no need to compare a biology teacher's 2.9 with an English teacher's 2.4. It's not a trustworthy comparison.

Since I'm suggesting a pile of data that a computer cannot reduce to a number, I'm implying that human beings must look at it and try to assess what it means. Teachers need discriminating feedback about particular practices and strengths and weaknesses, and we must premise the whole operation of collecting it on the following crucial principle: *in teaching, as in writing, it is possible to be good in very different ways.* A teacher might be warm or cold, organized or disorganized, easy or hard—and still be good.

The logistics of dealing with these data are not really so daunting. Indeed the very fact of not giving in to ranking or bottom-line verdicts leads naturally to the only kind of system we should trust: one that invokes some human judgment, not just arithmetical calculation. Thus we need a number of very small committees. Each would look at all the findings for only a few faculty members (probably inviting the teachers also to look at the data and to comment). For remember that we don't need this official evaluation system for every teacher every semester. If we ensure that teachers themselves gather informal feedback from students in every course every semester and from occasional visits by colleagues, this official judging mechanism need only be used every two to three semesters for untenured faculty members and every four to five semesters for those with tenure. Not an impossible job. And the point of gathering lots of these student perceptions and not summing them up into a grade is that they would lead to informed discussions of teaching. That's what we need.

Low-stakes evaluation probably does more to improve teaching than the official kind does.

I have not talked much about *faculty* evaluation of faculty members, because that is not my subject in this essay. But I assume that each evaluation committee would send a member to visit the classes of the teachers allotted to it, would get examples of the teachers' assignments and of their comments on student writing, and would also look at the teachers' grading (as well as probably getting evaluations from a sprinkling of former students). And—very important—the committees would get immeasurable help from seeing reflective statements by teachers about their strengths and weaknesses as teachers and about how they've developed since the last such statements. All these efforts will create faculty portfolios. Reviewing these portfolios is not such a difficult task if small committees only have to look at three or four teachers each. The job is also easier if committee members remember the problems of ranking and the value of judging. That is, the committees aren't trying to rank teachers with scores or to create precise bottom-line *verdicts* of how bad or good teachers are—with an important exception: the committees need to identify very poor teachers and very good ones. (And these end-of-spectrum verdicts are not so hard to agree on.) The committees' main goal is to analyze and communicate strengths and weaknesses for the vast majority of faculty members who are neither terrible nor extraordinary—so as to help these colleagues teach better.

I'll summarize my main points.

- It is impossible to have truly fair single-number rankings—that is, to get a range of good observers to agree in their verdicts about a complex performance. But if we do *less* evaluation, we can do it more carefully and thereby make it a bit more fair. We can avoid the simplification of ranking and instead use judgment—a process of careful looking that discriminates among features or dimensions of a complex performance and is built on the recognition that observers have different priorities. Thus students will be no better than we are at ranking (perhaps worse), but they are good at aiding informed judgment by giving us information on what the teacher did, what they learned, how they reacted.
- We can easily cut down on official, high-stakes, summative evaluation because we can get good results from more frequent low-stakes, informal, private, formative evaluation. Low-stakes evaluation probably does more to improve teaching than the official kind does.
- Though people are accustomed to ranking almost everything—looking for "bottom line," quantitative verdicts along a single continuum—we seldom need these oversimple verdicts. Yes, we need an unequivocal, blunt decision when someone's teaching is genuinely unsatisfactory or exemplary. But most of the time we are better off with more discriminating, multidimensional feedback about the strengths and weaknesses of particular features or practices.
- We must find ways to dignify student evaluations of teachers and to make the process thoughtful and reflective rather than mechanical.

Notes

[1] In particular, I found great relief when I realized that I don't have to grade individual papers just because I have to give grades at the

0605-b

end of the semester. On the papers I offer general responses and feedback about strengths and weaknesses and stop there, telling students that if they want to know how their final grade is shaping up, they can come see me starting at midsemester, but not sooner. See Belanoff and Dickson on portfolios.

[2] Someone is bound to object: "How can you say that teachers do not grade reliably in comparison with one another when some students get A's from all their teachers?" Reply: When a student gets all A's it's not an instance of the same performance getting the same ranking from multiple observers. Straight A students (typically people who care a lot about A's and know how to get them) have to make nontrivial adjustments in their performance from teacher to teacher. The performance that earns an A from one teacher will often enough *not* earn one from another without being adapted. Students give good testimony of how they must frequently make these adjustments. We see the same principle if we look at the other side of the coin: talented students who do *not* care about getting A's (and often they are brighter than the typical straight A student) usually get a fair number of B's—and not infrequently even some lower grades. The point is that such students *refuse* to adjust their performance from teacher to teacher.

Works Cited

Belanoff, Pat, and Marcia Dickson. *Portfolios: Process and Product.* Portsmouth: Boynton, 1991.

Diederich, Paul. *Measuring Growth in English.* Urbana: NCTE, 1974.

Elbow, Peter. "Visiting Pete Sinclair." *Embracing Contraries: Explorations in Learning and Teaching.* New York: Oxford UP, 1986. 179–97.

Gardner, Howard. *Frames of Mind: The Theory of Multiple Intelligences.* New York: Basic, 1983.

Kirschenbaum, Howard, Simon Sidney, and Rodney Napier. *Wad-Ja-Get? The Grading Game in American Education.* New York: Hart, 1971.

Lewis, C. S. *Studies in Words.* 2nd ed. Cambridge: Cambridge UP, 1967.

Smith, Barbara Herrnstein. *Contingencies of Value: Alternative Perspectives for Critical Theory.* Cambridge: Harvard UP, 1988.

Starch, Daniel, and Edward Elliott. "Reliability of Grading Work in History." *School Review* 21 (1913): 676–81.

———. "Reliability of Grading Work in Mathematics." *School Review* 21 (1913): 254–95.

———. "Reliability of the Grading of High School Work in English." *School Review* 20 (1913): 442–57.

Appendix

Countering Common Misbeliefs about Student Evaluations of Teaching

Robert Boice

What can be said in response to widespread beliefs that student evaluations of teaching (SETs) merit little credibility? I encourage colleagues to reconsider such attitudes toward SETs via four simple steps. The first consists of recognizing myths about SETs in their common forms. The second entails challenging [these myths] in light of the research literature. The third consists of inducing faculty to try SETs in formative and painless fashion to experience the value of feedback from students. The fourth helps show faculty how to educate students to give more constructive feedback in SETs.

First and Second Steps: Recognizing and Challenging Myths

The following myths about SETs appear most commonly in my experience. (Each myth is followed, parenthetically, with rebuttals from the research literature.)

1) *SETs reflect little more than a teacher's personality and popularity.* Some of us employ this belief to help salve the pain of evaluations. Statements often take this form: "If I were an entertainer, my SETs would improve dramatically." (In fact, the gist of research is that measures of personality and popularity correlate at low, usually insignificant, levels with SETs.)

2) *SETs mirror course difficulty and expected grades.* Here again, we can devalue SETs by assuming that they decrease as we make courses tougher. The common statement: "Highly-rated colleagues pander for good evaluations by giving easy assignments and generous grades." (Research makes a strong case to the contrary. Anticipated grading and SETs tend to be uncorrelated; heavy work loads correlate *positively* with SETs.)

3) *SETs of tough teachers improve when students are resurveyed years later.* All of us, as teachers or parents, like to believe that our "charges" will appreciate us more later, once they have seen the wisdom of our discipline. Some professors cite anecdotes to this effect to excuse their currently low SETs. (Research, sadly, shows that SETs remain remarkably stable over periods of many years. In other words, demanding and misunderstood teachers generally do not get higher ratings in retrospect.)

4) *Teaching is idiosyncratic and cannot be measured meaningfully.* This misbelief, that teaching defies analysis, is used to reject SETs because they supposedly miss the unique qualities of professors' styles. (In fact, research shows that effective teaching consists of rather ordinary and measurable factors like clear communication and rapport. Moreover, as we shall see anon, SETs lend themselves nicely to added measures that tap dimensions not covered in standardized forms.)

5) *Students are quick to complain and criticize.* If faced with low SETs we may suppose that students expect too much and disapprove too readily. (Studies of SETs, in contrast, suggest that students evaluate us generously; sometimes at rates of 80% for combined categories of good and excellent.)

6) *SETs reflect little more than classroom performance.* Some professors faced with disappointing SETs dismiss them because they cannot identify what they need to do differently. (When observations extend beyond the classroom, however, the problems may become apparent. A common example: acting in ways before and after class that students see as abrupt and impersonal.)

Third Step: Remedying Another Myth, That SETs Must Be Painful

As we begin to recognize that SETs may be credible, we may worry even more about the pain of getting poor evaluations from instruments that we now know are valid. One way of involving faculty in SETs that will be carefully considered and acted upon is to make the instrument painless. The first sample SET at the end of this article offers just such a format. By asking students simply to indicate desired directions of change along continua with no good or bad endpoints, faculty can get painless feedback about ways in

Appendix reprinted from *Teaching Excellence* 2.2 (1990–91) by permission of the Professional and Organization Development Network in Higher Education.

which they might consider change. At its best, the painless SET becomes the topic of discussion with classes (e.g., "why do you suppose that as many of the indications for change face in one direction as the other on this item?").

In my experience, once previously reluctant faculty try painless SETs, they are far more likely to volunteer for greater investment in conventional SETs.

Items on the painless SET can, of course, be changed to suit the tastes and needs of those who administer it.

Fourth Step: Countering a Final Myth, That SETs Must Come at the End

The obvious problem in not giving SETs earlier than at semester's end is that faculty are unlikely to make changes that could help improve ongoing classes. The second sample of an SET format at the end of this article illustrates a simple means of getting early, informal feedback from classes.

Early and informal evaluations like this one offer several advantages: a) They encourage faculty to rely on more than casual comments as the index of how they are doing. Instead, faculty can actively solicit anonymous opinions from all students—even those who ordinarily remain quiet during the semester. b) Early evaluations help get students involved. As the instructions attached to the early SET indicate, students can help collect, analyze, and even discuss the results. c) Discussions of the results of early SETs in class help educate students as evaluators. Faculty discussing early SETs can do more than indicate intended changes in teaching-related behaviors. They can also give students feedback on what kinds of evaluative comments are constructive and which are not. Experience with this strategy indicates that many students become more proficient as evaluators and more interested in the teaching process as a result of paying attention to specific categories of performance. d) Early SETs provide an opportunity to collect something usually left out of evaluations—compliments.

This general plan for getting faculty to abandon the temptation to see SETs as capricious indices of pandering and vengeful students revolves around action. It stimulates us and our colleagues to supplant our usual passiveness with proactiveness. In actual practice, I find that the general sequence of steps outlined here works best to change attitudes and behaviors (not necessarily in that order). In essence, these steps involve educating ourselves about what SETs really mean and how they can help.

One advantage of the sort of approach suggested here, according to my own research, is that it leads to three positive changes: 1) raised SETs, 2) alternative teaching behaviors, and 3) improved classroom comfort for both faculty and students.

State University of New York at Stony Brook

Representative References

Ellis, R. S. (1985). Ratings of teachers by their students should be used wisely or not at all. *Chronicle of Higher Education,* Nov. 20, p. 88.

Feldman, K. A. (1986). The perceived instructional effectiveness of college teachers as a function of their personality and attitudinal characteristics. *Research in Higher Education,* 24, 139–213.

Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by . . . *Research in Higher Education,* 30, 137–194.

Kulik, J. A., and McKeachie, W. J. (1975). The evaluation of teachers in higher education. In F. N. Kerlinger (Ed.), *Review of Research in Higher Education,* vol. 3. Itasca, IL: Peacock.

Menges, R. J., and Mathis, B. C. (1988). *Key Resources on Teaching, Learning, Curriculum and Faculty Development.* San Francisco: Jossey-Bass.

Weimer, M. (1990). *Improving College Teaching.* San Francisco: Jossey-Bass.

Student Feedback for Instructors

Bob Boice and Lyle R. Creamer

Recommend changes by drawing a directional arrow on each line. For example: +→+ Or use up arrow for no change: +↑++

SAMPLE ITEMS

Students should be less involved in class.	+-+-+-+-+-+-+	Students should be more involved in class.
Lectures should provide less detail.	+-+-+-+-+-+-+	Lectures should provide more detail.

Suggestions for Using the "Informal Student Evaluation" (ISE)

1. Administer the ISE at least once before formal evaluations; the earlier the administration of the ISE, the more instructors generally benefit. Try to use the ISE by midterm at the latest.

2. Allow 5 minutes at the end of a class to administer the ISE. Simply say that you're interested in learning what you're doing well and what you could do better while there is still time for change.

3. Ask for student volunteers to collect and compile evaluation sheets. In fact, students do not see this request as an imposition. In fact, students provide more useful feedback if they know that you will not see their handwriting (thus the reliance on students to collect and summarize the evaluation sheets).

4. Ask the student volunteers to summarize the results on a copy of the ISE. Numerical ratings can be summarized as a sampling of the most common types (e.g., "the instructor treats students with respect"). Have the summarizers omit uncommon remarks.

5. Xerox copies of the summary sheet and distribute them to all students at the beginning of the next class. Plan to spend 5 minutes reflecting on the results and probing students about what some evaluative comments mean (and how you can address them in terms of changes in style, content, etc.).

6. Use the occasion to educate students about ways to provide useful feedback to you; about your assessment of the class on dimensions like involvement, preparedness, etc.; and about your rationales for teaching the way you do (i.e., you may want to defend some of your practices).

7. Choose a sample of items from the formal evaluation to be used later in the semester (as in the example ISE provided here). These can give you a preliminary sense of how students will rate you (and a chance, in your discussions with them, to determine the basis for their numerical ratings on formal items).

Informal Student Evaluation (ISE)

1. What the instructor does well (please be specific):

2. What the instructor could do better (please be specific):

3. Please rate the instructor on the following scales 1–7 (7=maximum/excellent)

SAMPLE ITEMS

a. objectives and procedures were made clear. ____
b. instructor is well-prepared and organized. ____
c. the course stimulates my thinking. ____
d. presentations are clear. ____