8

EVALUATING: ASSESSING AND ENHANCING TEACHING QUALITY

> Reduce teaching to intellect and it becomes a cold abstraction; reduce it to emotions and it becomes narcissistic; reduce it to the spiritual and it loses its anchor to the world. . . . Good teaching cannot be reduced to technique; good teaching comes from the identity and integrity of the teacher.
>
> —Parker J. Palmer
> *The Courage to Teach* (1998, pp. 4, 10)

Lines on a vita can be counted to gauge research productivity, outside experts can provide commentary on the impact of one's scholarship, peers can attest to the critical contributions made in service, but how can one's teaching effectiveness be measured? Courses vary dramatically in size, level, and content, and professors' approaches to their classes run the gamut from formal lecture to open-ended discussion, structured to unstructured, cautious to risky, slow- to fast-paced, and practical to theoretical. Yet, most colleges systematically evaluate the quality of the instruction their professors provide by surveying the students themselves. These student evaluations of teaching, or SETs, if reliable and valid indicators of teaching, can be used to guide faculty development and personnel decisions. But if these inventories are used as the sole source of information about teaching and without considering their construct validity, then they can be transformed from useful resources into obstacles to overcome.

* * *

The professor's world is an evaluated one. Just to join it, one must pass a succession of tests, graded papers, and oral examinations that culminates in the defense of the dissertation. To land that first job in academia, prospective faculty members are interviewed, quizzed, and critiqued by search committees, deans, department chairs, and the faculty. The articles professors write, if published in the best journals, are often written,

reviewed, and revised again and again until they are eventually deemed acceptable. If professors are practicing psychologists, they must take and pass a licensing exam. Their grant proposals are sent to banks of experts who scrutinize their ideas before deciding if the proposals warrant further review, let alone funding.

This evaluative edge extends to the classes professors teach. Most colleges and universities evaluate the quality of instruction by regularly reviewing the adequacy of course offerings; tracking retention and graduation rates; and monitoring the quality of the library, technologies, and other resources students will use to reach their learning goals. Most universities also collect data about each teaching professor's competence in the classroom. In many cases the term's end turns the assessment tables on professors; they grade their students' learning with final exams, but the students grade their instructors' skills with "student evaluation of instruction" forms. Professors typically pass through a relatively detailed review each year when administrators make decisions about wage and salary increases, but the most elaborate evaluations are saved for promotion and tenure decisions.

All this evaluation is needed to sustain personal and professional standards. After all, psychologists are supposed to be "fully trained, keep up-to-date, and be good at what they do. Otherwise they should stop doing it" (Swenson, 1997, p. 64). *Personal evaluation* double checks individuals' subjective, and potentially biased, assessment of their adequacies. *Summative evaluation* serves the profession's and institution's purposes, for through evaluation they ensure that they live up to their obligations to serve students, parents, and the public. This evaluation also functions as feedback to professors as they refine their skills and extend their expertise; it is through practice paired with *formative evaluation* that the unskilled become skilled, novices become experts, and rookies become pros.

Few would argue against evaluation, in general, but the specifics of how, when, and for what purpose are more often points of critical debate. Many universities, for example, rely heavily on one particular source of information when evaluating faculty—students' ratings of their teachers' skills—and many faculty feel that these data are too distorted to be useful. The audience for the evaluation must also be considered when designing the feedback system, for the kind of information that will help instructors improve their teaching may be different from the kind of information that administrators need to make decisions about salary, promotion, and tenure. This chapter considers these issues, but interested readers may wish to also consult Braskamp and Ory (1994), Cashin (1995), and a November 1997 *Current Issues* section of the *American Psychologist* featuring papers by Greenwald and Gillmore, d'Apollonia and Abrami, Marsh and Roche, and McKeachie.

Dr. Greenwald was surprised when he opened the envelope that held the summary of his student ratings of his instruction for his undergraduate course in social psychology (Greenwald, 1997). He expected they would be good, because when he taught the course just the year before he got glowing marks from his students—the best evaluation he had ever earned. But this year's evaluations were not positive. In fact, they were the most negative reviews he had ever received. The shift was enormous—it spanned 2.5 standard deviations—even though he had used the same teaching and testing methods in the two classes. And he himself had not changed, had he? Wasn't he the same Dr. Greenwald who taught the course just the year before? Had his teaching skills suddenly eroded, or was the problem the source of the evaluation: Can students accurately judge their professors' instructional skills?

Most colleges systematically review the performance of their faculty, and each college's approach to this task is based on local norms, procedures, and historical precedent. But most colleges, despite their uniqueness, rely heavily on one key source of data: student evaluations of teaching, or SETs. These surveys yield a great deal of useful information about teaching, but many faculty question the meaning of the scores themselves. The most vehemently debated issue concerns the reliance on students' opinions and perceptions when evaluating teachers, but a number of related issues must also be considered, if only briefly: Are general impressions of teaching effectiveness more accurate than ratings of specific aspects of teaching? Does the use of student evaluations contribute to grade inflation? Should these ratings be used to make decisions about wages, tenure, and promotion?

Reliability of SETs

Reviews of the vast research literature dealing with SETs—estimated by Cashin (1995) at well over 1500 studies—generally agree that students' evaluations of a given instructor are reasonably stable across different rating forms, times (e.g., mid-term vs. end-of-term rating periods and immediately after class vs. delayed postclass follow-up), and courses taught in the same year. Test–retest reliabilities are high, as are internal consistency estimates of multi-item scales, even with scales with as few as five items (Marsh, 1982). Ratings also do not change much with the passage of time; when researchers tracked down students one year after they had completed the target course, they found that these retrospective ratings correlated .83 with the ratings students gave at the end of the course (Overall & Marsh, 1980). Interrater reliability is also high. Sixbury and Cashin (1995), for example, found that the median intraclass correlations across all items of a SET

survey ranged from a low of .69 for a 10-student class to a high of .91 for a 40-student class.

The Structure of SETs

Most SETs include global summary items and specific items. The global items ask students to rate the general quality of their instructor, the course, and their learning with such questions as "On a scale from 1 to 5, how would you rate this instructor?" The specific items focus narrowly on the elements of good teaching, including knowledge of the subject, enthusiasm for the material, respect displayed to students, and so on. The forms may also invite students to express their evaluation of the course in their own words with open-ended items such as "Why did you rate this instructor as you did?" In many cases, too, professors can select the items they wish to have included on the evaluations from a bank of items or devise their own unique queries about their teaching.

Do these items capture the meaning of effective teaching? Some investigators, noting the complex, multifaceted nature of the teaching process, have argued in favor of a complex, multidimensional inventory (Braskamp & Ory, 1994; Centra, 1993; Feldman, 1989). Marsh and Roche (1997), for example, argued against the use of global items because teaching is too complex and multifaceted to be measured with a single item such as "How effective was your instructor?" They based this conclusion on more than 2 decades of work by Marsh and his colleagues with the Students' Evaluations of Educational Quality (SEEQ) inventory of teaching effectiveness (e.g., Marsh, 1982, 1983, 1984; Marsh & Hocevar, 1991; Marsh & Roche, 1993). This measure asks students to rate specific characteristics of the class and the instructor—such as degree of organization, skill in stimulating discussion, rapport with students—but factor analysis of these items yields the following key components of effective teaching:

- *Organization of presentations and materials:* use of previews, summaries, clarity of objectives, ease of note taking, preparation of materials
- *Group interaction:* stimulating discussion, sharing idea/knowledge exchange, asking questions of individual students, asking questions to entire class
- *Breadth of coverage:* contrasting implications, conceptual level, and giving alternative points of view
- *Learning/value of the course:* challenge to students, value of material, amount of learning, increase in understanding
- *Rapport, or student–teacher relations:* friendliness toward students, accessibility, interest in students
- *Examinations/grading:* value of examination feedback, fairness of evaluation procedures, content-validity of tests

- *Instructor enthusiasm:* dynamism, energy, humor, style
- *Workload/difficulty:* perceptions of course difficulty, amount of work required, course pace, number of outside assignments
- *Assignments/readings:* educational value of texts, readings

Marsh and Roche reported that factor analyses of the SEEQ, with data collected in more than 50,000 classes including more than a million students, have confirmed the 9-factor structure of the inventory and their position on the multidimensionality of SETs. They wrote:

> Confusion about the validity and the effectiveness of SETs will continue as long as the various distinct components of students' ratings are treated as a single "puree" rather than as the "apples and oranges" that make up effective teaching. (Marsh & Roche, 1997, p. 1195)

Other investigators, in contrast, feel that the global items on the SETs are the more valid items—particularly when the evaluation will be used to make personnel decisions (e.g., Cashin, 1995; Centra, 1993; McMillan, Wergin, Forsyth, & Brown, 1987). McMillan et al. (1987) suggested that students' general perceptions of instructional effectiveness are more accurate than their perceptions of less "visible" aspects of teaching: for example, their ability to assess their professors' level of preparation, respect for students, or scholarly heft (Funder & Dobroth, 1987). Scriven (1981) argued that many professors are very successful instructors even though they do not score high on scales that measure enthusiasm, warmth, or organization. D'Apollonia and Abrami (1997), after reexamining the results of prior factor analytic studies of SETs, concluded that a single principle component accounts for 63% of the variance in SETs. They speculated that this teaching "g-factor" can be divided into subcomponents, including presentation, facilitation, and evaluation skills, but these subskills are all included in a General Instructional Skill factor (Abrami, d'Apollonia, & Rosenfield, 1997).

Construct Validity of SETs

Researchers and educational experts have yet to agree on a single indicator of the construct "teaching quality," so the definitive study of SETs has yet to be conducted. Researchers have, however, examined the relationship between SETs and a number of variables that should be related to teaching quality. They have found, for example, that SETs are generally highly correlated with the ratings of the instructor provided by other, presumably more objective, observers. For example, researchers have confirmed that SETs are significantly correlated with ratings provided by

- administrators (Feldman, 1989; Kulik & McKeachie, 1975),
- colleagues (Feldman, 1989; cf. Marsh & Roche, 1997),

- alumni (Braskamp & Ory, 1994),
- trained observers (Feldman, 1989),
- trained coders of specific instructional behaviors ("low inference" ratings, H. G. Murray, 1983), and
- faculty rating themselves (Feldman, 1989).

SETs are also related to student performance. Investigators have confirmed that the students in classes taught by professors who are more skilled —as indicated by their higher SETs—get better grades and higher scores on exams. Correlational studies of this grades–ratings relationship cannot rule out the possibility that these higher grades reflect the leniency of the professor rather than the professor's teaching skill. However, quasi-experimental studies have suggested that (a) some instructors teach better than others and (b) SETs are accurate indicators of who these teachers are. Researchers have carried out these multisection validity studies at colleges and universities that offer multiple sections of the same course. These sections, even though taught by different instructors, use the same syllabus, text, and—most important—the same final examination. Meta-analyses of dozens of these multisection studies generally confirm the relationship between students' scores on the final examination and instructors' ratings; students with better teachers—as identified through student evaluations —get higher grades on their finals (Abrami, Cohen, & d'Apollonia, 1988; Cohen, 1981; McCallum, 1984).

Bias in SETs

College courses vary dramatically in size, level, and content. Some require students to pore over original texts, some use elementary textbooks, and others use no book at all. Some meet at 8 a.m., others at 7 p.m. The professors teaching these courses vary in race, gender, age, experience, and so on. These factors do not affect the quality of the instruction, but do they not influence students' perceptions and ratings?

Table 8.1 summarizes some of the many and varied findings pertaining to the impact of extraneous factors—such as amount of work assigned, sex of instructor, and size of class—on SETs. As most faculty realize, factors beyond their control—such as the number of students in the class, their interest in the material prior to registering, the course level, and even the printed instructions on the evaluation form—are systematically related to ratings. But what faculty may not realize is that these factors account for only a small percentage of the variance in ratings. For example, one study indicated that teachers who voluntarily have their courses rated get better evaluations, but this statistically significant relationship accounted for less than 1% of the total variance in evaluations (Cashin & Perrin, 1983). Marsh (1980) suggested that extraneous factors taken together probably

TABLE 8.1
A Sampling of Research Conclusions From Studies of Biasing Factors in Student Evaluations of Teaching

Possible bias	Students' ratings of instruction
Academic discipline	Highest in humanities, lowest in hard sciences and math, and moderate for psychology and other social sciences (Cashin, 1990)
Administration procedures	Lower if anonymous and professor not present (Marsh & Dunkin, 1992); no difference if ratings taken at mid-semester or end of semester (Feldman, 1978)
Class size, time of day it meets	Higher in smaller classes than larger classes (Sixbury & Cashin, 1995); no effect of meeting time (Aleamoni, 1981)
Course level	Higher in graduate courses; some evidence that upper level courses higher than lower level, introductory courses (Aleamoni, 1981)
Description of rating's purpose	Higher if survey states that responses will influence tenure/promotion, salary decisions (Braskamp & Ory, 1994)
Grade	Higher if students get higher grades or expect to receive higher grades (Greenwald & Gillmore, 1997)
Instructor's research productivity	Higher if instructor is active in research (Feldman, 1987)
Instructor's age and rank	Slight tendency for younger faculty to receive higher ratings, but findings are inconsistent and differences are small (Feldman, 1983); graduate student instructors receive lower evaluations (Braskamp & Ory, 1994)
Instructor's sex	Men receive slightly higher ratings in simulated teaching settings, but women receive slightly higher ratings in field studies (Feldman, 1992, 1993)
Race	Insufficient data to draw conclusions (Centra, 1993)
Student motivation	Higher in elective courses; higher in courses that students rated as more interesting prior to enrolling (Marsh & Dunkin, 1992)
Workload and course difficulty	Higher in more difficult, demanding courses (Centra, 1993; Sixbury & Cashin, 1995)

account for only 12–14% of the variance in ratings. As Marsh and Roche (1997) concluded:

> Particularly for the more widely studied characteristics, some studies have found little or no relationship or even results opposite to those reported here. The size, or even the direction, of relations may vary considerably, depending on the particular component of students' ratings that is being considered. Few studies have found any of these characteristics to be correlated more than .30 with class-average students' ratings, and most relations are much smaller. (p. 1194)

Negative Effects of SETs

SETs are designed to assess how well professors are performing their duties in the classroom, but some analysts worry that SETs may have some deleterious effects. Surveys of faculty at various universities, for example, suggest that evaluations may undermine faculty morale, particularly among faculty with weak publication rates or a strong involvement in teaching (Armstrong, 1998). Because these faculty's careers are defined more by their teaching than by their research and service, a failure in this sphere will have more profound emotional and motivational consequences (Niedenthal, Setterlund, & Wherry, 1992). These professors may, for example, lose interest in teaching, particularly if the evaluations do not reflect the amount of time and energy they put into their teaching. As Armstrong (1998) wrote:

> Faculty members with poor ratings might decide that teaching is not rewarding and spend less time teaching. Teachers might get discouraged by ratings if they see no clear relationship between their attempts to provide a useful learning experience and their ratings. Teachers may get discouraged because time spent on teaching activities has little relationship to ratings or because, as they develop knowledge in the field through their research, there is no increase in their teacher ratings. (p. 1223)

McKeachie (1997) also worried that faculty may alter the way they teach to increase their ratings, but because students "prefer teaching that enables them to listen passively" professors may unwittingly adopt less effective, but more student-pleasing methods (p. 1219).

These evaluations may also contribute to grade inflation: the awarding of higher and higher grades for work of lower and lower quality. Of all the factors listed in Table 8.1, only students' expectations about their grades in the class is correlated with higher evaluations and under the control of the instructor. Some educators therefore fear that professors may be tempted to use grades to "buy" better ratings from students. They grade more leniently and lighten the workload for students, who reciprocate by giving them higher evaluations. Given the compressed nature of SET ratings on most campuses, if a lenient grading policy gains the instructor as little as a half-point on his or her average class evaluation, he or she may be catapulted from the lower third of teachers in the department to the upper third (Redding, 1998).

Faculty may not deliberately "dumb down" their courses to get higher evaluations. Instead, once-strict graders may unwittingly relax their standards as a result of pressure from students and administrators. Faced with complaints that their courses are too difficult and grades too low, they alter the way they test, the number of readings they assign, and reduce the

workload. The course becomes easier, grades rise, and so do SETs (Greenwald & Gillmore, 1997). On the other hand, they may grade leniently on purpose and cheat the system. Tabachnick et al. (1991), in their survey of teaching psychologists, found that only 40% felt that deliberately inflating grades was unethical, and a substantial proportion admitted they sometimes gave students better grades than they deserved just to "ensure popularity with students" (p. 510; see also Table 6.1 in this volume, chapter 6).

Even though SETs stand accused of fueling grade inflation, they may not deserve the blame. The *grading-leniency explanation* of the findings assumes that students appreciate receiving higher grades than they deserve and so they reciprocate by rating these kindly professors more favorably. But the *validity hypothesis* suggests that SETs and grades are correlated only because they are both caused by a third variable: the professor's superior teaching skills. The students in the class get better grades not because their professor is a lenient grader and the course is not demanding, but because the professor teaches so well that students learn more and so score higher on assessments. The *preexisting differences hypothesis*, on the other hand, suggests that students' prior interest in the course determines both their grade and their rating of the professor. Those students who are excited about learning psychology do well, and their excitement for the course raises the professor's rating, but students who are uninterested in psychology do less well and they do not give their professor high marks. It may be, too, that professors have a natural tendency to teach to the better students in their class, and these students therefore give their instructor higher ratings (McKeachie, 1997). Whereas Greenwald and Gillmore (1997), drawing on their structural equations modeling of the relationships between expected grade and course evaluation, recommended that statistical interventions should be taken to adjust SET ratings for leniency, McKeachie (1997), Marsh and Roche (1997), d'Apollonia and Abrami (1997) and other researchers in this area did not feel the findings are sufficiently strong to warrant this step.

Controversies and Convergences

What is the final word on SETs? Are they valid indicators of teaching effectiveness, or are they so easily manipulated by unprincipled professors that high ratings, like students' high grades, have lost their value? As Table 8.2 indicates, researchers have yet to reach complete consensus on matters of construct, convergent, and discriminant validity. d'Apollonia and Abrami (1997) and Greenwald and Gillmore (1997), for example, felt that ratings reflect students' general appraisal of their instructors, but Marsh and Roche (1997) felt that ratings, and the perceptions they measure, are differentiated and multidimensional. Whereas Greenwald and Gillmore (1997) concluded that ratings are substantially influenced by irrelevant

TABLE 8.2
Four Perspectives in the Debate Over the Validity of Student Evaluations of Teaching

Authors	Validity concerns and focal questions			
	Conceptual structure: Are ratings conceptually unidimensional or multidimensional?	Convergent validity: How well are ratings measures correlated with other indicators of effective teaching?	Discriminant validity: Are ratings influenced by variables unrelated to effective teaching?	Consequential validity: Are ratings results used in a fashion that is beneficial to the educational system?
Marsh & Roche	Like effective teaching, ratings are conceptually and empirically multidimensional. Their validity and particularly their usefulness as feedback are undermined by ignoring this multidimensionality.	Different dimensions of student ratings are consistently related to effective teaching criteria with which they are most logically related, thus supporting their construct validity.	Ratings are relatively unaffected by potential biases. Bias (mis)interpretations typically fail to control valid effects on teaching (e.g., class size, enthusiasm) that ratings accurately reflect.	Multidimensional ratings, augmented by consultation, improve teaching effectiveness (their most important purpose). Their use in personnel decisions, however, should be more informed and systematic.
d'Apollonia & Abrami	Although teaching is multidimensional, ratings contain a large global factor, which consists of several highly correlated lower order factors.	Global student ratings or a weighted average of specific ratings are moderately correlated with teacher-produced student learning.	There is little evidence of bias in ratings; few characteristics have been shown to differentially affect ratings and teacher-produced student learning.	Ratings provide valid information on instructor effectiveness. However, they should not be the only source of information, nor should they be over interpreted.

Greenwald & Gillmore	Because student ratings are dominated by a global evaluative factor, many ratings items detect only this global evaluation rather than their intended distinctive content.	Ratings measures show moderate correlations with achievement in multisection design.	The same instructor gets higher ratings when giving higher grades or teaching smaller classes. Older research indicates also that ratings are increased by enthusiastic style.	The quest for high ratings subtly induces lenient grading, which can both (a) reduce academic content of courses and (b) feed grade inflation.
McKeachie	There is a *g* factor in ratings, but there are also discriminable lower order factors.	Student ratings provide valid, albeit imperfect, measures of teaching effectiveness.	Influences on ratings by variables other than teaching effectiveness are of concern in the context of the deplorable practice of computing ratings averages that are compared with norms.	Ratings contribute to judgments of teaching effectiveness, but their use could be improved.

Note. From "Validity Concerns and Usefulness of Student Ratings of Instruction," by A. G. Greenwald, 1997, *American Psychologist, 52,* p. 1185. Copyright 1997 by the American Psychological Association. Adapted with permission.

factors, including the grades students expect to receive, d'Apollonia and Abrami (1997), Marsh and Roche (1997) and McKeachie (1997) felt that these biasing factors account for so little of the variance in ratings that they can be ignored with little risk. These latter investigators are more convinced that SETs provide a relatively accurate picture of a professor's classroom skills, but even they noted that SETs can be easily misinterpreted. They suggested that SETs are essential to the formative and summative review process, but as the next section notes, these ratings are only one source of information about teaching.

IMPROVING THE EVAULATION PROCESS

Teaching evaluation systems, like professors' systems for grading their students' performance, serve formative and summative functions. As formative reviews, they can provide specific, useful feedback about what does and does not work in the classroom. When the review is positive, the formative review inspires faculty to continue their good work, but when it is negative, it guides their personal development efforts. As summative reviews, evaluation systems provide evidence of the overall quality of the institution's effectiveness, and they also provide information relevant to administrative decisions on faculty hiring, salary, contract renewal, tenure, and promotion. Formative evaluations may help faculty improve their teaching skills, but summative evaluations provide the extrinsic motivation that translates the feedback into action.

Improving Formative Assessments

Formative assessments are more descriptive than evaluative, for they are designed to give instructors information about their success as lecturers, discussion leaders, testers, graders, and classroom managers. They do not yield grades or scores or rankings, but instead context-specific information about professors' progress toward their teaching goals. Professors who have not yet mastered all the intricacies of teaching should use these assessments to identify the factors that are blocking their progress. And even highly successful teachers should use formative assessments to check for unexpected problems in their teaching.

Student Rating Scales

Students, given their vantage point in the classroom and their familiarity with a variety of professors and their methods, are an excellent source of descriptive information about their professor's practices. Although, as noted previously, their perceptions are not in all cases 100%

veridical, when students' opinions about strengths and weaknesses converge, professors should take heed.

Because the more specific the feedback the better, global items such as "How effective is your instructor?" should be supplemented with items that ask about specific characteristics of the professor and class. The SEEQ, for example, collects students' judgments on a series of relatively specific items, such as, "You found the course intellectually challenging and stimulating," "Instructor's explanations were clear," and "Methods of evaluating student work were fair and appropriate" (Marsh, 1982, pp. 90–91). Other assessment systems, such as the Purdue Research Foundation's Cafeteria Course and Instructor Appraisal System, let faculty select the items they wish to have included on their assessment from a bank of over 140 items. Faculty may also want to develop their own list of items to include on a survey, particularly when they seek feedback about a particular nuance or innovation.

Open-Ended Verbal Descriptions

Many faculty feel that the most useful information they receive about their teaching comes from student responses to such items as, "Do you have any additional comments?" or "Please describe why you rated this course as you did" that are included in many assessment surveys. Even though Braskamp, Ory, and Pieper found that open-ended and fixed-response formats yield similar types of information, with correlations between these two types of measures ranging from .75 to .93 (Braskamp & Ory, 1994; Braskamp, Ory, & Pieper, 1981; Ory, Braskamp, & Pieper, 1980), the open-ended measures are more diagnostic—and painful—in some cases. To reduce the number of incomplete forms and increase the number of useful comments professors should

- explicitly ask students to add written comments,
- assure them that these comments will be read,
- administer the evaluation forms at the beginning of class rather than at the end, and
- code these responses rather than reviewing them haphazardly.

Individual and Group Feedback

Faculty, mindful that the official SETs will be administered at the end of the semester, sometimes overlook opportunities to assess their teaching earlier in the semester. Such assessments, because they can be gathered quickly and analyzed informally, provide useful information about the current class, and so may suggest changes that can be put into practice immediately.

The *midterm course check* procedure, for example, collects students' responses to the following three questions: What do you like the most

about this class? What do you like the least about this class? What one thing would you like to see changed? Or, as Angelo and Cross (1993) recommended, ask students to give examples of specific things that help them learn psychology, specific things that make learning more difficult, and practical suggestions for improving their learning. Students should be cautioned to not put their names on their comments, and also be reminded to try to focus on things that can be changed (e.g., amount of discussion, lecture style) rather than things that cannot be changed (when the class meets). Their comments can be categorized and discussed in a feedback session in the following class.

This approach can also be carried out as a collaborative group activity by asking a colleague to administer a Small Group Instructional Diagnosis. The colleague should separate the class into groups of five and give the groups about 20 minutes to answer the three questions in the previous paragraph. The groups must also select a recorder or spokesperson. Then, in a plenary session, the colleague pools all the ideas on the overhead and pushes the group toward consensus. Later, in a private meeting, the colleague relays the feedback to the instructor. This approach promotes collaboration and the development of consensus among class members on issues of classroom management and evaluation (Bennett, 1987).

Classroom Assessment Techniques (CATs)

Some of the most useful information about teaching effectiveness can be gathered by focusing on a particular aspect of the class rather than by seeking general information about overall course quality. A professor may, for example, wonder if students have too much time or too little time to complete the work assigned. Another may be worried that students' notes do not accurately reflect the contents of his lecture. Another may hope that students are learning to apply class material in their everyday lives, but be unable to assess her success in reaching this goal.

Angelo and Cross (1993) recommended using classroom assessment techniques, or CATs, to measure these specific instructional outcomes. They outlined a series of steps that faculty should follow in such assessments:

1. Select a course. Identify a single course that you will review using a CAT. This course should be one that you teach regularly, that you would like to improve in some way, but that has no glaring problems.
2. Identify a relatively specific teaching goal or question for this class. You can begin by reviewing the overall goals of the class, and narrowing down the focus of the review as much as possible. You may also want to think about portions of the class that usually do not go as well as you think they should,

and focus on that problem in your review. Angelo and Cross (1993) recommended that faculty begin by completing their Teaching Goals Inventory discussed in chapter 1. The professor may wish instead to begin with a specific question about some aspect of the class, such as why students are not interested in the material, what nonmajors and majors hope to get out of the course, or why students respond so negatively to classroom discussions.

3. Design an assessment method that will yield information about the question. The assessment should focus on what students have learned. As the instructions to contributors page in the journal *Teaching of Psychology* recommends, "empirical assessment should directly measure the impact of the technique on student learning (e.g., a pretest/posttest analysis of learning) rather than student self-report of learning" (Smith, 2001). The assessment may also include questions that will provide information needed to interpret the results, such as students' perceptions of problems, strengths, and weaknesses. Angelo and Cross (1993) and Table 8.3 describe several of these relatively simple, qualitative approaches to assessment.
4. Conduct the session that you wish to examine as you normally would. The assessment intervention, because it focuses on student learning, should fit naturally into the session's teaching and provide students with feedback about goals.
5. Carry out the assessment procedure, being certain that students understand that the intervention is not a test of their learning, but an indication of the adequacy of the lesson. Angelo and Cross recommended giving students credit for participating, but also keeping responses anonymous.
6. Analyze the data. The professor should review the responses generally, perhaps by reading them over in a single session to get a general sense of their contents. The data can then be reviewed more thoroughly by taking counts on the number of students who missed specific types of material or voiced similar concerns about the class. Specific cases should also be culled to use as illustrative examples of strengths, weaknesses, and areas for improvement.
7. Draw conclusions based on the data. If the results are unexpected or inconsistent, spend some time mulling them over, discussing them with students in the class individually, or sharing them with colleagues. Consider such general questions as, "Do your data indicate how well (or poorly) students achieved the teaching/learning goal or task?" and "Can you

TABLE 8.3
Examples of Classroom Assessment Techniques Discussed by Angelo and Cross (1993)

Assessment technique	Objective checked	Description
Empty outlines	Accuracy and depth of student's notes	At the end of the day's lecture the professor gives students a sheet of paper with only the major headings of the lecture listed, and asks students to fill in subheadings and key points
Memory matrix	Students' ability to compare and contrast concepts	Students complete a table that lists concepts or theories down the rows and their characteristics across the columns (e.g., classical and operant condition are row entries and terms such as "reinforcement," "extinction," and "shaping" are column headings)
Minute paper	Grasp of key points of presentation	Students are given 1 minute to identify the points that they feel were the most important ones in the day's presentation and ask questions they want answered
Muddiest point	Identification of areas of uncertainty	Students are asked to identify the area of the lesson that was the muddiest, or least clear, to them
One-sentence summary	Students' ability to integrate information	Students must write a grammatically correct single sentence that summarizes a topic; one variation asks students to answer the questions "Who does what to whom, when, where, how, and why" in one sentence
Application cards	Student's ability to apply course material to new examples	The professor hands out large index cards to students, who are asked to write down at least one application of the day's presentation to a real-world situation or problem

interpret why you got the results you did?" (Angelo & Cross, 1993, p. 55).

8. Give the students feedback about the assessment. The results of the assessment can be communicated with students through a didactic session where the professor covers the findings and offers interpretations or by the preparation of a more formal report that is distributed to students. If the re-

sults suggest changes in method, these possible changes should be discussed carefully with students and, depending on the specificity of the syllabus, initiated the next time the course is taught.

9. Evaluate the assessment. Angelo and Cross recommended reviewing the effectiveness of the assessment procedures, noting any ways that the intervention could be improved to yield clearer, more interpretable information.

Consider, as an example, the use of classroom assessment by a professor who teaches a course in learning and cognition. His assessment is triggered by his suspicion that students are not connecting the course content to problems that they face in the own lives. After explaining the reasons for the exercise with the class, he puts this question on the overhead (Angelo & Cross, 1993, p. 68):

> Have you tried to apply anything you learned in this unit on human learning to your own life? Yes or No.
>
> If "yes," please give as many specific, detailed examples of your applications as possible.
>
> If "no," please explain briefly why you have not tried to apply what you learned in this unit.

Students were asked to use a word processor to generate a one-page response, and the paper's due date was set for the next class period. Students were not to put their names on the papers, but the professor noted who turned in a paper and gave each student credit.

One professor who used this method reported that 60% of his students claimed they were using the course's content to improve their studying methods, enhance their memory, reduce their stress, deal with their children's behavior, and so on. At the next class he reviewed the findings with students, and with the class developed a more detailed listing of possible applications. The professor now stresses applications as a specific goal in this course, and conducts the assessment regularly to check his teaching effectiveness.

Collaboration With Colleagues

Colleagues can be an excellent source of formative feedback. Informally, they can act as a sounding board for new ideas, a supportive audience to listen to difficulties, and an advisor who can recommend solutions. More formally, they can review the materials of the course—syllabi, tests, lecture notes, Web materials, and so on—and identify strengths, weaknesses, and revisions. They can also visit the classroom itself, and write up the results of their visit in a report or share them with the instructor over a cup of coffee. Faculty observers, however, tend to be lenient reviewers, and one

colleague's high appraisal of a learning strategy might not be shared by another colleague down the hall (Centra, 1975). One may therefore wish to consider providing a checklist for observers to use to structure their comments. Murray (1983) and Mintzes (1979) described observational inventories that are less vulnerable to observer bias because they focus on discrete, specific types of behavior. These low-inference ratings ask observers to indicate only the extent to which the professor displayed behaviors that are related to effective and ineffective teaching, such as speaking clearly and expressively, smiling or laughing, using concrete examples, using headings and subheadings, showing interest in the subject, showing concern for students, and so on (Murray, 1983, pp. 140–141).

Improving Summative Assessment

In teaching, as in all things, many paths lead to excellence. One professor may be a superb lecturer who teaches students so stealthily they do not even realize their neural networks are being rewoven. Another may be the quintessential discussion leader who can draw out and organize students' viewpoints in a rich texture of insights. Others may develop novel methods of instruction, write textbooks that inspire students, or mentor colleagues in the craft of teaching. The steps taken in evaluating teaching must reflect this diversity. Because professors reach excellence through many different paths, no single index or indicator of quality fairly captures this diversity in style and substance. A summative review should take into account not only professors' classroom teaching but also the caliber of the instructional and evaluative materials they develop and use in their classes, the academic quality of the course's contents, the quantity and quality of their nonclassroom teaching activities, and their overall contributions to the discipline's educational mission.

Classroom Teaching

What is the best source of information about professors' competence in the classroom itself: their skill when lecturing, when leading discussions, and when answering questions; their ability to motivate students to learn the material; their work in a teaching laboratory; or their effectiveness as tutors when discussing recent empirical findings with advanced students? As noted earlier in the chapter, studies of the validity of student ratings of their teachers' effectiveness, although not entirely consistent in their conclusions, suggest that summative evaluators should solicit students' opinions rather than rely on their own. Annual reviews of faculty, tenure and promotion decisions, and considerations for wage increase, if they are at least partially based on the quality of professor's teaching, should therefore consider what students say about what goes on when their professor is teaching. Specific suggestions include:

- Although some assessment experts recommend providing reviewers with detailed information about specific facets of teaching, most favor the use of a small number of items that require a general, overall evaluation of teaching. Rather than asking, for example, about skill in lecturing, leading discussions, enthusiasm, building rapport, or providing feedback, summative rating items such as, "How would you rate this instructor's overall teaching effectiveness?" and "Rate this course on a scale from 1 (very poor) to 5 (outstanding)" are preferable. These items are general enough to be asked in any class, no matter what its size, procedures, or level, yet they are highly correlated with other indices of student learning (cf. d'Apollonia & Abrami, 1997; Marsh & Roche, 1997).
- SETs should be used to generate only overall ratings of faculty's teaching—for example, *exceptional*, *meets standards*, or *unacceptable*—rather fine-grained, multicategory discriminations (d'Apollonia & Abrami, 1997). This conservative approach prevents reviewers from reading too much into the numbers and reaching conclusions that are not warranted given the possibility of measurement error. Moreover, as McKeachie (1997) noted, in most cases

  > personnel committees do not need to make finer distinctions. The most critical decision requires only two categories—"promote" or "do not promote." Even decisions about merit increases require no more than a few categories, for example, "deserves a merit increase," "deserves an average pay increase," or "needs help to improve." (p. 1218)

- Because SETs are survey data, they should be discounted if their validity is threatened by unusual administrative procedures and inadequate sample sizes. Cashin (1995) recommended that evaluations should be based on at least five sections, taught in different years. SETs should be interpreted cautiously in classes with fewer than 10 students, and if a substantial portion of the class (30%) did not complete the forms.
- SET information should also include data pertaining to grade expectations, grade distribution, and student motivation; scores can also be statistically adjusted to control for these influences (Greenwald & Gillmore, 1997).
- If merit pay and promotions are based, in part, on teaching effectiveness, then SETs should be administered in all classes, and the same generic questions should be used on all surveys. The use of standardized items promotes the development of

norms pertaining to teaching, but only if all faculty are required to have students complete evaluations: professors should not have the option of not evaluating their instruction.

The Quality of Instructional and Evaluative Materials

Outstanding teachers, in addition to stimulating learning through direct instructional activities, also teach by developing effective instructional materials, activities, assignments, and assessment methods. They may not be mesmerizing presenters or skilled discussion leaders, but they can teach effectively with well-designed Web sites, by giving students detailed feedback about their individual work and by setting clear classroom goals and providing students with the resources they need to achieve them. The quality of these procedures will likely be indicated by students' evaluations of the course itself, rather than their rating of the instructor. Colleagues can also review these instructional materials. As Centra (1975) noted, colleagues are too inaccurate for use as classroom observers. They are, however, excellent judges of instructional material and course management. Just as faculty are skilled in reviewing a scholarly article and determining its publishability, faculty are capable of reviewing a colleague's teaching materials to determine if they are excellent, adequate, or need improvement. Instructors can facilitate this process, however, by preparing a dossier, or portfolio, that describes their teaching methods, their educational philosophy, and includes copies of material used in classes (e.g., syllabi, tests, handouts, classroom exercises, sample lecture notes, graded examinations). This important element in summative evaluation in teaching is examined in detail in chapter 9.

The Academic Quality of the Course

There are good courses in psychology, but there are also great courses. One professor may cover all the topics when teaching introductory psychology and measure students' performance adequately, but another may challenge students to think critically about the field's key issues, coordinate a series of student-generated research studies, provide students with opportunities to express their understanding of psychology in their own writing, and have time left over to help students apply psychology in their everyday lives. Summative evaluations should attempt to gauge the relative academic quality of the course itself by looking past *how* the class is taught to focus on *what* is being taught. In most cases, members of the professor's own academic unit can judge whether a course meets the discipline's standards for academic quality by asking such questions as:

- Is the course material current?
- Is the instructor adequately trained in the subject that he or she is teaching?
- Is the course pitched too low, in that it is so easy that students who learn very little nonetheless pass it?
- Does it cover the material that the college catalog says it is supposed to cover, or has it wandered from its purpose to focus on trivia?
- Is the course intellectually challenging?

The Quantity and Quality of Nonclassroom Teaching Activities

When summative evaluators base their ratings of faculty only on classroom teaching, they unwittingly endorse the view of those who criticize faculty for spending too little time teaching. Yet much teaching occurs outside of classroom settings, through the following indicators:

- *Advising and mentoring:* the number of advisees; participation as advisor on undergraduate thesis, graduate thesis, and dissertation committees; any reports (both favorable and unfavorable) from advisees pertaining to advising.
- *Publications dealing with teaching in higher education:* (a) papers and texts published or presented on educational topics, (b) manuals developed for classroom use, (c) papers published or presented with student-coauthors (both graduate and undergraduate), and (d) textbooks.
- *Specialized teaching:* nonclassroom-based teaching, such as (a) public teaching (presentations to the community at large, including speeches, workshops, educational newspaper articles, and interviews); (b) individualized instruction, including mentoring and tutoring; (c) workshops for colleagues and advanced students; (d) distance education; (e) interdisciplinary teaching.
- *Curriculum development activities:* description of courses developed or substantially changed. Innovations in teaching courses or topics should also be noted.
- *Service contributions in teaching:* administrative duties or service that focuses primarily on teaching, such as participation on any departmental, college, or university committees and task forces dealing with teaching.
- *Supervision and mentoring:* guiding students' work on individual research projects, thesis and dissertation research, the development of clinical skills, and other forms of graduate teaching.

Overall Contributions to the Discipline's Educational Mission

Ideal professors do all things well. They teach in the classroom, on the sidewalk, in their offices, through technology, with dramatic effect. Whether they are lecturing, leading discussions, questioning, or mentoring, their students learn. But ideal professors reach beyond fine teaching, per se: They make broader contributions to teaching practices in their disciplines and to higher education in general. Such contributions as research into pedagogical practices, curricular reform, university- and national-level service in teaching, public teaching, and mentorship of other teachers dot the vitae of the finest teachers. They are concerned with their own and others' teaching, to the point that they study the process and hone their own skills. They participate in formal and informal analyses of teaching not because they are experts, but because they are always seeking improvement.

Evaluating Evaluation

Faculty evaluations, whether conducted to help faculty improve their teaching or for input into personnel decisions, should be conducted with care. Formative reviews can provide professors with suggestions on how to improve their teaching, but not if the evaluations themselves are invalid —or thought to be invalid. Summative evaluations, too, must be based on more than a simplistic bean count of faculty's gold stars given them by their students. Summative evaluators who factor teaching skill into their reviews of faculty are to be commended for not basing merit awards only on research productivity, but if they base their review on incomplete data, their good intentions will be for naught. Faculty should be evaluated, but these reviews must be based on procedures that are consistent the current state of knowledge in the field of teaching evaluation rather than the personal predilections of faculty or administrators.