A Short Guide to

Evaluating Teaching

written by

Elena Berman, Ph.D.

Assessment and Faculty Development Specialist
Assessment and Enrollment Research (AER)
University of Arizona

MLK Building, Rm. 200
(520) 626-4214
eberman@u.arizona.edu

Revised August 2002 and March 2003

A Short Guide to Evaluating Teaching

Revised August 2002 and March 2003

Acknowledgments/Disclaimer

This Short Guide sums up ideas from numerous evaluation experts. Readers are encouraged to review primary sources, cited in the References section (pp. 27-28). Michael Theall of Youngstown State University provided detailed comments about Chapters 1 and 2, resulting in a greatly improved text. Betty Atwater, Homer Pettey, Cecile McKee and others provided valuable suggestions for making this document more user-friendly. Most particularly, this Guide is influenced by Jennifer Franklin, who guided my reading, answered my questions, and generally helped me arrive at the view of evaluation that informs this Guide. Appendices A, B, and E have evolved from documents co-written with her as part of documentation for AER services or products.

Chapter 1. Overview

The climate for evaluating teaching has changed greatly over the past three decades and continues to be in a state of ferment. Partly in response to external constituencies (legislatures, boards of regents, funding and accrediting agencies, the public), universities at all levels are acknowledging the need for better practice in evaluating teaching. Demands for post-tenure review have sparked considerable discussion of the importance of fair evaluation, as have challenges to the practice of tenure itself.

This increased attention to evaluation goes hand in hand with a more general shift in the way evaluation is being viewed. In the past, evaluation was often regarded as a sort of add-on. Students learned, teachers taught, programs proceeded, and at the end, ways were found to arrive at grades, scores, success indicators. The emerging view makes evaluation more central to the processes of teaching, learning, and program development, in the sense of first defining goals and standards, and second, providing the feedback necessary to attain those goals and standards. In this view, evaluation becomes an ongoing part of the processes of teaching, learning, and program development, shaping these activities in directions deemed desirable.

As a consultant at an evaluation service center, I am often asked **how** to evaluate teaching. That question cannot be meaningfully answered unless we know first **what** we're evaluating, and second, **why** we're evaluating. Only after we know what we want to scrutinize, and why, does it make sense to reflect on how to go about it.

Purposes for Evaluation

There are two basic purposes for evaluating anything, typically called "formative" and "summative" (Scriven, 1967). Decisions about sources of evaluative data, methods of collecting information, and the importance of a formal process all hinge on whether the primary purpose is summative or formative.

Summative evaluation is evaluation to judge results. For faculty, summative evaluation occurs in the context of tenure and promotion, merit raise, and hire/fire decisions. Because it affects people's livelihood, a high standard of fairness is important. For example, policies and procedures must be formally stated and uniformly applied.

Formative evaluation is evaluation for improvement or development. In the case of teaching, it may refer to activities faculty engage in to develop and improve as teachers. (These may range from reviewing a videotape of oneself in the classroom to studying a machine-generated item analysis report of test results to using short feedback questionnaires relating to understanding course content or appreciating classroom activities.) "Formative evaluation" may also refer to review by a chair or administrator done solely for the purpose of prompting improvement. Formative evaluation is typically individualized, self-determined, and informal.

Summative and formative approaches may seem potentially at cross-purposes. In formative evaluation, the critical question is "Where are the weak points?" While it is important to know what is going well, the emphasis is on finding out what isn't working in order to make changes. The primary summative question, "How good is it?" might seem in conflict with an approach that looks for problems. However, when evaluation is viewed as an ongoing process, it is clear that the conflict is only apparent. Formative evaluation done consistently over time should result in improvement with each iteration. From this perspective, early discouraging results only add luster to later success, while demonstrating the value of evaluation for improvement.

Interest in systematic summative evaluation of teaching has grown over the past 30 years, along with increased use of and reliance on student ratings results. Beyond student ratings, there are few time-tested, tried-and-true approaches. In fact, active experimentation is the norm both locally and nationally, with great differences both across institutions and within them. This is not surprising. Since responsibility for summative evaluation is primarily vested in departments, differences reflect disciplinary tradition, leadership, and priorities. Since an important predictor for success of a summative evaluation system is department buy-in, it is appropriate that departments develop and "own" their own systems.

In developing summative systems, methods traditionally used in a formative spirit have sometimes been converted to summative use without adequate attention to the special demands of summative review. Student ratings, for decades considered feedback that individual faculty could do with as they chose, are suddenly being invoked to deny merit raises or promotions. Where once faculty could ignore ratings statistics and focus on written comments, it is now important to understand the statistics and how they will be used. It is equally important that academic units specify precisely what the expectations are. Other evaluation methods developed largely in formative contexts in the United States, notably peer observation and teaching portfolios, are becoming requirements in administrative processes, often without clear standards for review.

While there are many differences between formative and summative purposes, an effective overall approach fosters connections on several levels. Summative considerations in part dictate the thrust of formative efforts. Data collected/created by the activities of formative evaluation become part of documentation submitted for summative review. Summative findings result in recommendations about formative processes. Most importantly, to meet the standards and requirements a department sets for summative evaluation, faculty should have access to numerous methods of formative evaluation to improve their teaching.

Focuses for Evaluation

Teaching, by its nature, cannot be deemed effective unless significant learning takes place. Therefore, the evaluation of teaching may logically focus either on teacher attributes/behaviors, learner outcomes, or both. Intuitively, it makes sense for student learning to be the primary measure of teaching effectiveness, and many educators like the idea of relying on learning outcomes rather than student or colleague judgments of presentation style or course materials. Certainly, a demonstration that significant student learning has taken place is a prima facie demonstration that teaching has been effective.

However, using learning outcomes to judge an individual's teaching ability is highly problematical. For example, while course grades are in some sense a measure of student learning, using course grades to measure teaching effectiveness would likely accelerate grade inflation. In any case, if the percentage of As, Bs, etc. is roughly uniform across courses in a department at a given level (as it often is), there is no way to distinguish among faculty on this ground. More fundamentally, grading practices are not consistent, neither within departments nor across departments. In some cases, grading clearly reflects student achievement, but in many cases, it does not.

Another possibility is relying on results from standardized tests, which, in some disciplinary areas, could provide a reliable measure of student achievement. However, standardized tests are unsuitable for many disciplines, and many faculty members are resistant to using them as a measure of student learning. Also, since such tests rarely measure the effects of a single course, it would be hard to attribute success to any particular instructor.

In any case, judgments of efficacy of teaching are only possible if inputs are measured. Imagine two classes that score identically on a standardized test. However, students in class A started out knowing half the material and were motivated to succeed while class B students started out underprepared and resistant

to instruction. In such a case, despite the convergence of results, it would be incorrect to conclude that instructors A and B have equivalent teaching ability.

Not only previous knowledge, but also attitudes and motivation influence student learning. Many college students already have considerable efficacy as learners and would probably learn regardless of the teacher, especially if they are internally motivated. While effective teachers can usually improve student motivation, motivation is impacted by a number of factors outside their control. It would seem unfair to penalize instructors whose students are poorly motivated or had other priorities.

Because of the difficulties of measuring student outcomes, the evaluation of teaching has tended to center on teacher attributes and behaviors. This makes sense, since teachers can alter their own behavior more readily than they can anyone else's. Considerable research has shown that good teaching can be broken down into a small number of factors (these are discussed in Chapter 5), and that most teaching skills are learnable.

If we base the evaluation of teaching largely on teacher attributes/behaviors, it would be nice to find that teachers rated highly using such measures are also teachers whose students are learning a lot. Research suggests that this is true. In studies involving sections of the same course taught by different instructors, with a common, independently graded final, students who scored highest on the final also tended to rate their instructors highest (Cohen 1987).

While summative evaluation of teaching cannot rely heavily on demonstrations of student learning outcomes, using student outcomes as feedback for improving teaching is a traditional practice of effective instructors. Documentation of efforts aimed at improving student outcomes can be a particularly convincing part of the documentation of effective teaching.

Chapter 2 of this guide addresses the formative evaluation of teaching. Chapters 3-5 deal with summative review from the perspective of first, the instructor being evaluated, and second, the academic department. Chapter 6 suggests ways of integrating summative and formative evaluation and offers guidelines for department practice, while Chapter 7 explores the importance of reviewing teaching at the unit level as well as the individual instructor level.

Chapter 2: Formative Evaluation of Teaching

Formative evaluation generally involves getting feedback early and often, and using it to good purpose. What aspects of teaching should be focused on? From the instructor's point of view, four (interrelated) areas stand out: 1) presentation/delivery, 2) course design, 3) class climate (i.e., accessibility, interaction, perceived fairness), and 4) student outcomes.

The Teaching Portfolio

A common expression in pedagogic circles these days is "reflective practice." The idea is that if you think about what you're doing as a teacher, you're likely to do it better. While some great teachers operate on charisma and instinct with little reflection, for most teachers, reflecting about teaching will pay off in improved teaching. (Of course, most "natural born teachers" also reflect on the practice of teaching.) In the United States, the "teaching portfolio" approach was widely seen as a method for fostering reflection. Those who developed teaching portfolios generally agreed that they resulted in new ideas as well as epiphanies about what worked and why some things didn't work. Portfolios were often developed in workshop settings where interactive activities among colleagues contributed to producing renewed commitment to and creativity about teaching.

"Teaching Portfolios" are now part of the official language of summative review at the University of Arizona, and are partially defined by university-level policy guidelines and partly by department guidelines (see Chapter 3). In some units, they more resemble the traditional dossier than the portfolio conceived as a product of reflection. Regardless, conducting formative evaluation in the reflective spirit originally associated with the teaching portfolio concept is recommended, both for its immediate effect of improving perceived teaching effectiveness (both by you and by students) and its long-term goal of providing data for use in summative evaluation. Moreover, the processes of getting formative feedback generate a variety of documents which can be included in a teaching portfolio to illustrate particular aspects of effective teaching.

1) Assessing presentation/delivery skills

Presentation/delivery refers primarily to effectiveness in the classroom. Since students see the instructor day after day and week after week, and are the intended audience for the presentations, they are an obvious source of assessment data about presentation skills.

End-of-semester student ratings questionnaires tend to emphasize aspects of presentation/delivery and are generally a good barometer of effectiveness in this regard. However, they offer little detail about the specifics of excellent or lackluster performance. For more information, the following approaches are recommended:

- enhanced student ratings questionnaires
- midsemester focus group evaluations
- classroom assessment
- peer/expert observation
- videotaping (self-observation)

Enhanced Student Rating Questionnaires

The UA student ratings system (TCEs) is designed to meet both summative and formative purposes. Instructors have the choice of a standard "short form" TCE questionnaire or one of a variety of "long form" questionnaires customized for particular class formats (e.g. lecture, discussion, lab, studio).

The standard Short Form Questionnaire contains eleven highly general questions suitable for use in summative review. The various long-form questionnaires contain diagnostic questions about behaviors strongly associated with effective teaching in particular contexts. These questionnaires draw upon extensive research regarding observable behaviors strongly correlated with teaching effectiveness (e.g. H.G. Murray, 1991).

Questionnaires may be seen at http://aer.arizona.edu/AER/teaching/questionnares/ques_main.htm.

Midsemester focus group evaluations

First developed at the University of Washington, mid-semester focus group evaluations (originally called SGID, for Small Group Instructional Diagnosis) provide a safe environment for students while giving the instructor immediately usable feedback related to a particular class. A facilitator meets with a class for 20-40 minutes (depending on class size), usually at the end of a class, and collects student feedback in response to two questions:

1. What are the strengths of this course?
2. What are your suggestions for changes?

A written summary of the results is sent to the faculty member, usually followed by a meeting to discuss the results and possible follow-up actions.

Midsemester focus group evaluations are currently supported by most colleges at UA. AER trains graduate students as focus group facilitators and supervises their work. Contact AER at *midsem@u.arizona.edu* for more information. Instructors can also enlist a colleague or graduate student to serve as facilitator, but to ensure a safe environment for students, the person chosen should not be able to identify individual students in the class. An outline of the AER procedure is available at http://aer.arizona.edu/AER/Teaching/midsem/fgdirections.htm

Classroom Assessment

A variety of activities can be used in class to collect feedback from students. *Classroom Assessment Techniques* (Cross and Angelo, 1993) provides numerous examples from a wide variety of disciplines categorized under three headings: "Assessing Course-Related Knowledge and Skills," "Assessing Learner Attitudes, Values, and Self-Awareness," and "Assessing Learner Reactions to Instruction." Strategies range from instructor-designed questionnaires to student-generated self-assessments. Some are widely known and require little set-up, such as the "Minute Paper," in which students briefly respond to such questions as "What's the most important thing you've learned from today's lecture?" or "What's the most important question you're left with after today's lecture?" Others are more complex, demanding more time both for preparation and interpretation.

Cross and Angelo's research shows that responses are frequently unexpected, prompting faculty "to rethink and redesign their teaching" (p. 372). Classroom research also promotes greater student involvement in learning.

Resources for Classroom Assessment

Thomas A. Angelo and K. Patricia Cross, *Classroom Assessment Technique.* Second Edition, San Francisco: Jossey-Bass Publishers, 1993.

Classroom Assessment Techniques: http://www.ntlf.com/html/lib/bib/assess.htm.

An Introduction to Classroom Assessment Techniques: http://www.psu.edu/idp_celt/CATs.html

Classroom Assessment Technique Examples: http://www.hcc.hawaii.edu/intranet/committees/FacDevCom/guidebk/teachtip/assess-2.htm

Classroom Assessment Techniques in the Sciences: http://www.flaguide.org/

Peer/Expert Observation

While observation by colleagues requires caveats if results are to be used for summative evaluation (see Appendix D - "Using Peer Observation in Summative Evaluation"), peer observation is an excellent method for improving teaching. It is especially effective when done reciprocally as part of a teaching circle or mentoring relationship. In fact, it could be argued that observing teaching improves the teaching of the observer as much as the teaching of the observee.

The advantages of formative peer observation are: 1) a sympathetic observer who can offer content-related as well as other feedback, and who can be directed to watch for particular behaviors and dynamics, 2) a gateway to faculty discussion of common instructional problems; increased collegiality within a department, and 3) a way of keeping faculty in touch with the experience of being a student.

Peer observation for formative purposes begins with a conversation about what the goals are for the class to be observed and what the instructor wishes feedback about. It is often helpful to use a guide or protocol for recording observations (see resources). In fact, a well-designed observation guide constitutes a mini-lesson in what is observable about effective teaching.

Peer observation may involve a general, overall look at in-class performance/dynamics or a more focused look at particular aspects of classroom dynamics. In one type of observation offered by the University Teaching Center, the observer maps which students speak up in class and what types of

responses they get. These maps eloquently reveal patterns of speaking to one side of the room or responding differentially to genders or ethnic groups. Call UTC at 621-7788 for more information.

When peers observe colleagues within the same discipline, they sometimes find it difficult to separate presentation considerations from judgments about content. If feedback is desired on aspects of teaching unrelated to content, consider being observed by a colleague outside your disciplinary area or by an instructional expert. Departments (or individuals) wishing to set up formative peer observation programs may consult AER for guidelines and sample protocols, which may also be found in the references cited below.

Resources on the Peer Observation Process

Brinko, K. T. and R. J. Menges, eds. *Practically Speaking: A Sourcebook for Instructional Consultants in Higher Education*. Stillwater, Okla: New Forums Press, Inc., 1997.

DeZure, Deborah, "Evaluating Teaching through Peer Classroom Observation," in Peter Seldin and Associates, *Changing Practices in Evaluating Teaching*, Bolton, Mass.: Anker Publishing Company, 1999.

Hutchings, Pat. *Making Teaching Community Property: A Menu for Peer Collaboration and Peer Review*. American Association Higher Education (AAHE), 1996.

Keig, Larry and Michael D. Waggoner. *Collaborative Peer Review: The Role of Faculty in Improving College Teaching*. ASHE-ERIC Higher Education Report No. 2. Washington, DC: The George Washington University, School of Education and Human Development, 1994.

Weimer, M., J.L. Parrett, and M. Kerns, *How Am I Teaching: Forms and Activities for Acquiring Instructional Input*. Madison, WI: Magna Publications, Inc., 1988.

Peer Observation of Teaching: http://www.ltsn.ac.uk/genericcentre/index.asp?id=17849

Videotaping (Self-observation)

Having yourself videotaped and watching the videotape with a consultant is arguably the most powerful method of getting direct information about how you come across to a class and what the classroom dynamics are. Even for faculty who use active learning methods that minimize their "stage presence," watching a videotape of a class, including student workgroup time, can give information that's easy to miss when the class is actually taking place. A consultant can help you see strengths as well as weaknesses, offer suggestions for improvement, and provide information about resources. Studies show that faculty who review their videotapes with consultants are more likely to benefit than those who view them on their own (Brinko, 1997).

The University Teaching Center (621-7788) will videotape classes on request and review the video with the instructor.

2) Assessing course design skills

Weaknesses in course design tend to be more subtle than weaknesses in presentation/delivery. They are sometimes the problem when students complain that they are unprepared for or uninterested in a course, or when they (or you) sense that class time isn't being well spent. While students are usually aware in a general way when there are problems in course design, they are rarely able to provide direct feedback except about superficial aspects, such as pacing and sequencing.

A good course design integrates learners (students), course content, and instructional activities in mutually supportive ways with the goal of enhancing student learning. Some characteristics of good course design are listed below, along with suggestions for assessing the extent to which they characterize your courses. Many aspects of course design can be self-assessed by instructors; in some cases, peers and experts can offer important insights. Well-designed classroom assessment approaches (see above) are especially helpful. For the most part, effective course design is invisible to direct observation.

Characteristics of effective courses

course has clear purpose within the overall curriculum

To see how well a course articulates with other courses, use a pretest[1] to find out if you are beginning where students left off in previous courses. If there is a discrepancy, adjust your course content or work with instructors in feeder courses to adjust theirs. Ask instructors in successor courses to use pretests as a measure of whether students coming from your classes are adequately prepared. Find out how the material in your course relates to courses taken concurrently as well as sequentially.

amount of material is appropriate to allotted time and student level

Students will certainly notice if there is too much content relative to amount of course credit; they are also likely to notice if there is too little. Usually, the scope of coverage is adjusted over the first few offerings of a course until it seems about right. Colleagues are often able to make good determinations of whether the amount of material is appropriate. Classroom assessment and midsemester focus group evaluations can also provide relevant information. Pretests facilitate a determination of whether the content is appropriate to the level of the students.

contents are successfully related to student abilities and interests

With regard to abilities, performance on tests and assignments is a generally reliable guide. If too many students are getting high or low grades, then course content and assessments should be adjusted. Use pretests and classroom assessment methods to ensure that contents are related to student interests. Excellent examples may be found in Cross and Angelo (1993) and Diamond (1998).

appropriate instructional strategies are used for particular learning tasks

Consultation with instructional specialists can introduce you to new strategies and help you gauge which are likely to be successful. Cooperative learning, the case method, problem-based learning, self-paced learning – all are approaches that have proved effective for particular types of content and learning tasks. Reviewing pedagogical materials in your disciplinary area can lead you to approaches known to work for the kinds of content you teach. Most disciplines have a journal devoted to pedagogy in that discipline.

Closely related to instructional strategies are instructional media. Both the choice of media (appropriate to learners and learning tasks) and how effectively they are used may be assessed through surveys and other classroom assessment approaches.

assignment and testing plan is appropriate and keyed to course goals

Using a matrix to map course goals against assessment measures is a good way to check whether you're testing what you want students to learn (Jacobs and Chase, 1992). Research shows that there is frequently a discrepancy between desired learning outcomes and testing practices. Since there is strong evidence that students learn what they are tested on, it is good practice to make sure that you're testing the most important things you want students to learn.

class time is used effectively both on the course level and at each class meeting

Since effectiveness in this regard, to a considerable extent, is in the eye of the beholder, it makes sense to ask students themselves what they view as the best uses of class time, and whether they feel

---

[1] A pretest should survey student knowledge of the course content as well as prerequisite material. It is usual (and advisable) to cover student attitudes and expectations as well. See "Using Pretests" (http://aer.arizona.edu/AER/Teaching/docs/UsingPretests.PDF) for more information.

that topics, assignments, and exams are reasonably spaced across the semester. Midsemester focus group evaluations as well as classroom assessment approaches can provide relevant data.

Resources for assessing course design

Thomas A. Angelo and K. Patricia Cross, *Classroom Assessment Techniques.* Second Edition, San Francisco: Jossey-Bass Publishers, 1993.

Robert Diamond, *Designing and Assessing Courses and Curricula in Higher Education*. San Francisco: Jossey-Bass Publishers, 1998.

3) Assessing class climate (interaction, accessibility, perceived fairness)

Do students feel free to speak in your classroom, and are they comfortable questioning you via email or during office hours? The most straightforward measures of accessibility and interaction are how much students interact, and how many times they seek you out. Counting student visits and email contacts can be used as such a measure, as well as data obtained from classroom assessment approaches. Written comments on TCE forms often speak to these questions, and they are usually brought up during midsemester focus group evaluations, both as strengths and as weaknesses.

Perceived fairness is another strong measure of class climate. When students feel they are not being treated fairly, they may stop participating in other ways. (Some may stop attending.) Students will usually complain if they feel the grading is unfair; they may write comments to this effect on ratings questionnaires and they almost always bring this up during midsemester evaluations. Instructors can also directly query students about perceived fairness (cf. Ory and Ryan, 1993, pp. 128-30). One UA instructor includes on each test a section that invites students to "rate the test" – asking about best and worst questions as well as about overall coverage.

Resources for assessing class climate

Ory, J.C. and K.E. Ryan, *Tips for Improving Testing and Grading.* Newbury Park, CA: Sage Publications, Inc., 1993.

4) Assessing Student Learning Outcomes

To evaluate whether you're doing a good job teaching, it's a good idea to find out what and how much students are learning. This is the broad and complex area of testing and grading, the subject of a forthcoming *Short Guide*. Four excellent books on the topic are listed below.

Resources for assessing student learning outcomes

Anderson, R.S. and B.W. Speck. *Changing the Way We Grade Student Performance: Classroom Assessment and the New Learning Paradigm*. San Francisco: Jossey-Bass Publishers. New Directions for Teaching and Learning, No. 74, Summer 1998.

Jacobs, L.C. and C.I.Chase. *Developing and Using Tests Effectively: A Guide for Faculty*. San Francisco: Jossey-Bass Publishers, 1992.

Ory, J.C. and K.E. Ryan. *Tips for Improving Testing and Grading.* Newbury Park, CA: Sage Publications, Inc., 1993.

Walvoord, B.E. and V.J. Anderson. *Effective Grading: A Tool for Learning and Assessment*. San Francisco: Jossey-Bass Publishers, 1998.

Chapter 3. Summative Evaluation of Teaching – for Faculty

The summative evaluation of teaching has two faces, depending on one's perspective. For academic units and departments, the challenge is to develop a valid, reliable, flexible, and practical system of evaluation. For those being evaluated – individual faculty members wishing to compile appropriate documentation of effectiveness – the primary tasks are to teach well and to understand department policy and practice. Suggestions for individual faculty are presented in this chapter, while Chapters 4-5 review the characteristics of effective systems and offer guidance for developing them.

Faculty members at the University of Arizona are required to demonstrate teaching competence at every stage of review. (See http://w3.arizona.edu/~vprovacf/ for university policy.) However, what this means varies from college to college and department to department. Given that both internal and external pressures are prompting changes in the activities of teaching itself (e.g. distributed learning, team-based approaches) as well as the standards for evaluation, it is likely that approaches to evaluating teaching will be in flux for some time.

Current department practice may be more or less explicitly stated. Both the amount of attention paid to teaching and the level of guidance on documenting teaching vary considerably. Faculty members should become familiar with practice in their department and seek mentors within the department who will provide reliable guidance. At the same time, faculty should recognize that department evaluation plans are likely to change, and that today's loose guidelines may be clearly articulated requirements tomorrow. It makes sense to be prepared to document instruction-related effort.

Experts recommend that faculty keep records of instruction-related activities, saving items that may later serve as documentation of teaching effort and effectiveness. This could include examples of student work, lists of textbooks reviewed, descriptive rationales for instructional choices, results of midsemester evaluations and classroom assessments, etc. Keep a separate file for each course that includes information about development, implementation, maintenance, and results. These files hold the source material for your Teaching Portfolio (see below). In addition to formal portfolio requirements, faculty submitting documentation for summative review may take the opportunity to present their own best case based on additional documentation they have collected.

Summarizing TCE Results

The Arizona Board of Regents (ABOR) has mandated that feedback from students be used in evaluating faculty. Teacher/Course Evaluations (TCEs) are the most widely used instrument for collecting this feedback at UA. Promotion and Tenure Guidelines require a "quantitative summary" of TCE results. AER provides a variety of reports that fulfill this requirement, notably the TCE History, the Overall Teaching Effectiveness Graphics, and the Comparison Summaries for each course. For more information, see our *Guide to Student Ratings at UA* (http://aer.arizona.edu/AER/teaching/Guide/TCEGuide.pdf). UA instructors may access their own reports through the Individual Instructor Reports button on the AER website.

Summarizing Student Written Comments

Using student written comments in summative evaluation is controversial and problematical. Evaluation experts agree that precautions must be taken to ensure validity, reliability, and confidentiality. See Appendix B ("Using Student Written Comments in Summative Evaluation") for discussion.

Selecting other materials for submission: The Teaching Portfolio

As noted earlier, teaching portfolios have more commonly been used formatively than summatively in the United States. The summative use of portfolios raises many questions about consistency and reliability that will need to be addressed at the unit level. (See chapters 4-5.)

Generally speaking, a teaching portfolio is a collection of documents demonstrating commitment and excellence in teaching. A characteristic feature of portfolios as described in the literature (see Resources, below) is an interpretive narrative explaining the portfolio contents along with reflections on teaching strengths, interests, and areas of projected development.

According to Peter Seldin, the teaching portfolio enables faculty to document "both the complexity and individuality of good teaching." Achievements relating to course or curricular innovations, participation in faculty development efforts, and special services for students can be highlighted; differences between faculty members and for the same faculty member over time can be accommodated.

Materials included in the teaching portfolio should highlight teaching strengths, focusing on how unit, college, or university goals are served by your teaching activities. Some documentation of performance in all the domains of teaching may be presented (see Chapter 5), with emphasis on one or more areas depending on unit and personal goals. Reviewers will appreciate brevity and careful organization.

Lists of items that may be included in a teaching portfolio have been presented in many places. Appendix C ("Possible Items for the Teaching Portfolio") offers a list sorted according to what dimension(s) of teaching particular documents highlight.

Resources on the teaching portfolio and other aspects of preparing for summative review of teaching

Diamond, Robert M. *Preparing for Promotion and Tenure Review: A Faculty Guide.* Bolton, MA: Anker Publishing Company, Inc., 1995.

Murray, John P. *Successful Faculty Development and Evaluation: The Complete Teaching Portfolio.* ASHE-ERIC Higher Education Report No. 8. Washington, D.C.: The George Washington University, Graduate School of Education and Human Development. 1995.

Seldin, Peter. *The Teaching Portfolio, A practical guide to improved performance and promotion/tenure decisions*. Bolton, Mass: Anker Publishing company, Inc., 1991.

Items in Portfolios: http://www.lgu.ac.uk/deliberations/portfolios/ICED_workshop/seldin_book.html

Chapter 4. Summative Evaluation of Teaching — for Academic Units and Administrators

Current Practice and the Impetus for Change

Summative evaluation occurs first and foremost within departments. In most departments, it has tended to be quite informal. The evaluator may be the department head alone (especially in the case of second and fourth year reviews), a standing committee, or a group convened strictly for this purpose. Typically, the evaluators each examine the materials presented by the candidate and base their judgments on their personal interpretations of the material. Rarely have there been clearly defined standards against which submitted documentation is measured. In fact, discussion about evaluating teaching has centered more on what material should be included in the packet than on how that material would be judged.

Because teaching was taken for granted, it was usually easy to agree that someone's teaching was good enough, extra good, or needed improvement. If it was at least good enough, the best case for it would generally be made, and that would be the end of it. Faculty who put considerable effort into teaching often felt dismayed that there was so little departmental reward or recognition for it. Others felt discouraged from putting extra effort into teaching due to the same negative reward structure.

As for teaching that needed improvement, if the instructor in question was well-liked and successful in research, poor teaching would often be glossed over. A common sentiment is that some wonderful researchers and colleagues are just not gifted as teachers, and really would do better if they didn't have to teach at all. Since required teaching seemed punishment enough for these individuals, there was a tendency to avoid "punishing" them further by requiring that they put more effort into teaching. (Pretenure faculty were often urged **not** to put too much effort into teaching, letting it suffice for teaching to be just good enough.)

Since in many cases, there have been few consequences for teaching poorly, few rewards for teaching well, and virtually no litigation over whether teaching evaluation was fair or appropriate, the informal approach worked well enough and "if it's not broke, why fix it?" As in most cases, the impetus to fix arises in three circumstances:

1. **External Pressure**: Accrediting agencies, boards of regents, or granting agencies insist on seeing clear standards and rigorously applied procedures.
2. **Internal Pressure**: It becomes obvious that the system is "broke" because either 1) someone litigates, arguing that he/she wasn't treated fairly, judged according to clear standards, given adequate time to improve or come up to standard, etc.; 2) it is clear to all that standards have been "bent" to keep/not keep someone about whom there was dissension, the department is polarized, and it is clear that the informality of the evaluation system was a contributing factor; or 3) there is clear agreement that a senior faculty member is doing a terrible job teaching, but there are no possible sanctions because it would involve invoking standards not consistently held to for others or ever before for him/her.
3. **Practical and Proactive Foresight**: Departments see the value of being forearmed against internal and external pressures, and agree that those who significantly further the department's teaching mission should be rewarded. They also see an effective evaluation system as an ultimate time-saver and a flexible instrument that will enable the department to improve both individual and collective teaching, make it easier to determine fair workload and accountability for faculty, and ultimately simplify program evaluation and result in improved strategic planning at the department level.

External pressures are usually the weakest motivators because the course of least resistance is to appease external constituencies with minimal action. Internal pressures are in some ways the strongest motivators, but at great cost. The third impetus is obviously the most favorable because a department moved by proactive foresight will not be in a hurry or at cross-purposes with itself. Since developing a fair and effective evaluation system requires an investment of time at the outset and ongoing attention, it is far more likely to be successful if the third impetus is operative.

Internal pressures are more likely to arise as the evaluation of teaching is taken more seriously and decisions have greater consequence. Judgments perceived as arbitrary have a chilling effect on department morale, beyond any consequences for individuals. When decisions are made without clear standards, or merit funds allocated without distinction between excellence and adequacy, or individuals warned about their teaching feel that they are being singled out unfairly, or new faculty are unsure what work will be valued, ultimately a whole department suffers.

Evaluation judgments made without explicit criteria and standards are frequently perceived as arbitrary. Research supports such perceptions. Many studies show that when colleagues judge peers based on intuitive understandings of effective teaching, they differ greatly in how they assign worth and criticism. When observing classes, they are particularly divergent in the conclusions they draw (see Appendix D ["Using Peer Observation in Summative Evaluation"] for a discussion of the summative use of peer evaluation based on observation). Colleagues also may differ in how they interpret student ratings, with some viewing excellence as the 4.0-4.5 range, while others consider the 4.5-5.0 range as excellent. Many lack a meaningful frame of reference or even vocabulary for evaluating syllabi, course plans, or websites. A good evaluation system clarifies this frame of reference and provides this vocabulary. As such, in itself it can act as a force for change. For example, if faculty know that a website will be judged for particular features, they will be more likely to incorporate those features into their websites. Good evaluation practice, both summative and formative, can greatly contribute to improved teaching and overall department excellence.

Characteristics of Effective Summative Evaluation Systems

In developing an effective summative evaluation system, it is helpful to know the characteristics of effective systems. These are:

- validity and comprehensiveness
- reliability
- explicitness, publicness
- flexibility
- time and cost effectiveness
- periodic self-evaluation
- support at the highest relevant level of administration
- linkage to formative evaluation
- linkage to planned change strategies

Validity and Comprehensiveness

An evaluation system is valid to the extent that it measures what it is intended to measure. Since the goal is to measure effective teaching, a fair summative evaluation system must define the parameters of teaching as well as set criteria for each parameter.

The question of "what" teaching comprises for the purpose of summative evaluation is far from trivial, since teaching is something of a moving target. If effective teaching today requires websites, webliographies, multimedia presentations, and orchestration of student project teams, these areas of competence and performance should be evaluated. While there is considerable consensus about the parameters of effectiveness for lecture/discussion teaching formats,[2] it is not clear how these relate to new instructional formats.

In an effective system for evaluating teaching, the full range of teaching responsibilities and activities is taken into account. When faculty complain that student ratings measure popularity rather than

---

[2] While to some extent, what constitutes effective teaching is particular, i.e. related to these students in this subject matter, researchers have converged on a small number of characteristics across wide disparities of topic and method. These are content knowledge and appropriate choice of goals; effective organization and exposition; ability to challenge and engage students. See Doyle (1984) for discussion.

genuine teaching effectiveness, they are raising a validity issue. Indeed, an evaluation system that relies solely on student ratings would fail to address such critical areas of teaching competence as content expertise and instructional design skills. Using only student ratings, which center on presentation/delivery skills and, to some extent, rapport with students, is like evaluating a chef based only on his/her ability to prepare appealing looking platters, without considering abilities like devising menus, balancing components of meals, preparing entrees, desserts, etc. A person with excellent presentation/delivery skills could receive an outstanding evaluation even if he/she taught out-of-date content or tested students at an inappropriate level of challenge.

Reliability

For a system to be reliable, it must work consistently from individual to individual and from iteration to iteration. This means that the criteria must be sufficiently general to accommodate a variety of teaching-related activities and teaching styles, and that evaluators interpret the criteria and standards similarly. If two evaluators judging classroom presentation differ in what they consider effective, so that evaluator A rewards effective lecturers while evaluator B gives the highest ranking only to those who involve students in active learning, evaluation is not consistent. Those being evaluated could rightly argue that candidates are being held to different standards. If evaluator A sees student ratings in the 3.5 to 4.0 range as mediocre, and evaluator B interprets the same numbers as satisfactory, the same problem obtains.

The keys to ensuring reliability are: 1) explicit statements of criteria, and 2) training evaluators to apply the criteria. Both are necessary for a reliable system.

Reporting formats that work across course types and can be easily and consistently interpreted go a long way to making reliability attainable. The challenge is to develop criteria and standards that accommodate a wide range of teaching approaches. Student ratings questions like: “Rate the effectiveness of the in-class activities,” and “Rate the effectiveness of out-of-class assignments” are designed to be appropriate for a wide variety of instructional approaches, yet allow a consistent metric across candidates. A question like, “Rate the instructor’s ability to facilitate teamwork” has much more limited applicability, yet a department wishing to foster team activities in its courses may wish to include such a question.

Explicitness/Publicness

The procedures and criteria should be known both to evaluators and those being evaluated. Just as students need to know the basis for grading and what they will be held responsible for, faculty need to know how they will be evaluated so they can prepare themselves successfully.

Flexibility

An effective system accommodates individual differences and changes over time for the same individual. Flexibility can be built into the system at numerous points: e.g. differential weightings of domains being evaluated, a variety of materials allowed as documentation, special credits for particular activities, an official appeal process.

Time and Cost Effectiveness

An evaluation system will not work if users perceive it to involve more time/trouble than it is worth. This means effecting economies everywhere possible (e.g. standardized forms, documents that can be retrieved from websites by administrative assistants) and frequently reminding people of the benefits of the system. When individuals see that good work is valued and evaluation results are positively linked to change, they are more likely to value time spent on evaluation.

Periodic Self-Evaluation

A good evaluation system is periodically monitored for time and cost effectiveness, as well as whether evaluation goals are being met and evaluation results being used productively.

Support by Administration

Without administrative support at the highest relevant level, it is unlikely that evaluation will be more than perfunctory. This means that department heads who wish faculty to take evaluation seriously must create an atmosphere in which evaluation is valued and supported. (Chapter 6 offers suggestions.)

Linkage to Formative Evaluation

If no resources exist to help faculty reach and exceed the criteria and standards required by an evaluation system, it is fundamentally unfair for faculty to be evaluated by those standards. For example, a department that required faculty to develop course web sites should make sure that faculty have access to appropriate training and support. Indeed, consultation and assistance should be available to faculty wishing to meet requirements at the highest level of achievement. Constructive links to formative evaluation can be supported at the department level through teaching circles, mentoring relationships, and other methods.

Linkage to Planned Change Strategies

An evaluation system that stops when candidates get their "final grade" misses the mark. If evaluation data is not used to guide change, perhaps the largest potential benefit of evaluation is missed. Without linkage to change strategies, evaluation is not likely to be time/cost effective.

Chapter 5 Key Steps in Developing an Effective Department Plan

A full list of steps for developing a unit plan for evaluating teaching is provided in Appendix E ("Developing an Academic Unit Program for Evaluating Teaching: A Checklist"). This chapter focuses on the central steps of deciding on the areas of teaching that will be evaluated, specifying required and optional documentation for each area, setting weights, and establishing criteria by which the documentation for each area is judged. Two further steps, norming (to ensure convergent interpretation of criteria) and integrating results, are briefly discussed.

An effective evaluation plan must be effective in two regards: first, it must be valid, reliable, etc., and second, there must be department buy-in and support. This chapter addresses the first of these issues; department buy-in is discussed in Chapter 6.

University of Arizona guidelines require a five-point evaluation system, with four points to be "overall satisfactory" and one to be "unsatisfactory." The names of the categories are left to be decided by departments; they may be called "outstanding, excellent, satisfactory, needs improvement, unsatisfactory," or "outstanding, satisfactory, fair, poor, unsatisfactory," or "A, B, C, D, F."

The challenge is to develop a process in which those being evaluated can be reliably and unambiguously assigned to the appropriate category. Recognizing that some instructors are clearly more effective than others, and also that in some departments, most instructors are genuinely outstanding, how can one clearly state a fair basis for making these determinations?

The five-point scheme forces a large degree of explicitness on the part of an evaluation system because it is important that each point be clearly interpretable. Specifically, an evaluation system should make it easy to answer questions like the following:

- If instructor A is deemed "outstanding," and instructor B is deemed merely "excellent," what is the basis for this distinction?
- If instructor F is deemed "unsatisfactory," is it clear that his/her teaching is measurably worse than that of instructor D, deemed "needs improvement"?

In thinking about developing a system for evaluating teaching that will make it possible to answer such questions, a useful analogy is the task of devising a fair grading system for a course. At the end, everyone will receive a "final grade" based on scores assigned to a variety of documents and activities. The documents may be as different as multiple choice exams vs. complex projects. They provide evidence that the goals of the course are met. In the case of teaching, documents may include a summary of TCE results; colleague judgments based on observation or examination of course materials; a video of teaching performance; a course portfolio including syllabus, study guides, exams, and analyses of exam results, etc. As in many classes, those being evaluated are a group of individuals with differing baselines, commitment, and motivation. A fair system must accommodate these differences while ensuring that all who "pass" meet reasonable standards and that those who excel are rewarded.

In evaluating students, it is generally agreed that standards must be stated at the outset and uniformly applied. Instructors should be accorded the same courtesies. Where there are multiple "graders," evaluees have the right to expect that all will use the same metric. Evaluees have the right to expect that "tests" will relate to stated expectations and purposes, and that they're taking the same "tests" as their peers.

As with course grades, the evaluation system is a powerful method for fostering the programs and initiatives that best support a department's strategic goals. A department wishing to encourage team teaching, course websites, or inclusion of undergraduates in faculty research activities can provide an incentive for faculty to engage in the desired activity by making it a factor in its evaluation system.

The approach presented below is hardly the only one possible. However, all approaches meeting the criteria described in the previous chapter will have many features in common.

Selecting areas to be evaluated and sources of data for each area

Thinking about what to evaluate takes place at two levels: 1) determining the general areas of teaching competence to be evaluated (e.g. presentation skills, course design skills, use of new technology/methodology, etc.), and 2) naming the documents/sources that will provide data for judging that area. These levels are often confounded in practice, but it is helpful to keep them distinct. Problems can arise when a department jumps directly to designating documentation without careful consideration of what is being documented. For example, a department might decide to require a summary of student ratings results, summaries of peer observation reports, and a videotape to be reviewed by the evaluating committee. All of these documents focus largely on presentation skills. In this department, an excellent lecturer will likely be favored over someone less inspirational, even if the latter individual has done a better job at course design, resulting in better-educated students. The choice of these documents informs faculty implicitly that what matters above all in this department is skill at delivery. While it is not always obvious what sorts of documentation are appropriate for other areas of teaching, beginning by naming documents makes it easy to favor areas easily documented and to overlook areas where an obvious method of documentation is lacking.

Various breakouts of the teaching role are possible. Arreola (1995) lists instructional delivery skills, instructional design skills, content expertise, and course management as one possible set of areas. In some departments, advising/mentoring might be added; in others, outreach or distributed learning activities. Ultimately, decisions should depend on department values and priorities. For example, if a department decided that teamwork skills were important for its majors, they could base a component of the evaluation of teaching on the extent to which faculty integrate teamwork into their courses. Credit could be given just

for using team-based approaches, with extra credit for evidence that team activities fostered exceptional student learning. By establishing that experimenting with teamwork can only have positive evaluation consequences, a department makes it more likely that faculty will experiment with teamwork.[3]

Once the areas to be evaluated are chosen, sources of evaluation data must be selected. (These are equivalent to the tests, assignments, and projects that form the components of the final course grade.) The more important the area being evaluated, the more desirable it is to have more than one source of data for that area. For example, if presentation/delivery skills count for 50% of the total score for teaching, relying on student ratings alone places a disproportionate weight on student opinion. Reliability will increase if student evaluations are supplemented with peer evaluations. A discrepancy between peer and student evaluations should trigger further inquiry.

While logically it makes sense to begin by determining the areas to be evaluated, and then decide on measures for evaluating each area, these processes tend to go hand in hand. For a quick validity check, list the domains of teaching being evaluated, then check to ensure that evaluation documents cover the territory. The matrix in Table 1 shows how this might look.

Table 1. Matching Sources of Evaluative Data with Domains of Teaching Activity

Domain of teaching / Eval document	Presentation/ Delivery	Course Design	Course Maintenance	Content Expertise	Faculty Development
TCE results	Y	N	N	N	N
Descry. teaching rel. activities	N	maybe	maybe	maybe	Y
course mats (peer-rev)	N	Y	maybe	Y	maybe
video (peer-rev)	Y	N	N	maybe	maybe
documentation of fac dev	maybe	maybe	N	N	Y

The matrix shows that presentation/delivery is evaluated based on two sources of data, while other areas have more limited documentation unless special provisions are made to ensure that data is included. For example, peer review of course materials may specifically document faculty development if course materials are supplemented with descriptions of efforts made in developing them. Based on this chart, course maintenance may not be adequately addressed by the documents listed.

[3] Bernstein (1996) describes a similar approach to the use of peer observation reports, apparently very successful.

Setting weights for each area of teaching and source of evaluation data

Once the domains of teaching and documents are chosen, weights must be specified both for overall domains and for specific documents. Different weighting schemas may be appropriate for different types of teaching assignment. Table 2 shows both a uniform weighting schema and a variable one for domains of teaching. The variable approach is one way of building flexibility into the evaluation system, with faculty negotiating percentages from year to year.

Table 2. Sample Weighting Schema for Various Dimensions of Teaching

Dimensions of Teaching to be Evaluated	Uniform Approach	Variable Approach*
Presentation/Delivery	50%	40-60%
Course design	30%	20-40%
Use of technology in teaching	10%	5-15%
Advising/Mentoring of students	10%	5-15%

*If a variable approach is chosen, percentages should be specified in advance for each candidate. Candidates should not be able to decide at the last minute that some dimension will be considered at the minimum or maximum percentage.

The next step is to assign weights to the various sources of data for each dimension. Explicit decision rules should spell out how each piece of documentation is weighted. Table 3 illustrates a possible assignment of weights for documentation of effective presentation/delivery.

Table 3. Example of Decision Rules for Assignment of Weights for Documents supporting effective teaching

Evaluation Data / Dimension of teaching	*Total Weight**	Student Ratings	Peer-reviewed Video	Other**
Presentation/Delivery)	*50%*	30%	15%	5%

* When combined with other dimensions of teaching, the final score for presentation/delivery will have a weight of 50%. Thus, student ratings will account for 30% of the final score.

**A variety of optional entries could be allowed.

Describing criteria for submission and criteria for judgment of each data source

For each dimension of teaching being evaluated, both acceptable documentation and criteria of judgment must be specified. For example, according to Table 3, presentation/delivery skills are to be documented by TCE results, a video (to be evaluated by a peer committee), and an optional component (possibly a report by a specialist or department head, summary of in-class evaluation, etc.). Tables 4a and 4b show sample statements of criteria of submission and criteria of judgment for videotapes. A sample statement of criteria of judgment for TCE results is provided in Appendix A ("Evaluating TCE Results").

Table 4a. Sample Statement of Criteria for Submission of Videotapes
a. may consist of a single class or a montage of edited "classroom events"
b. total time does not exceed 30 minutes
c. is accompanied by a written or spoken narrative explaining when and where the tape was made, and what the contents are intended to show

4b. Sample Criteria of Judgment for Videotapes	
Outstanding 5 pts.	Presentation demonstrates good organization, engaging delivery style, student involvement, constructive responses to student comments or orchestration of student discussion
Excellent 4 pts.	Presentation reasonably well organized, delivery style polished, appropriate student participation/orchestration of student discussion
Acceptable 3 pts.	Coverage of content is adequate, but cues to aid student understanding are deficient (e.g. failure to define new concepts, lack of examples)
Needs improvement 2 pts.	Content coverage is perfunctory, fails to engage students, is poorly organized, engages in substantial digressions, fails to orchestrate student discussion, etc.
Unacceptable 1 pt	Students not treated with respect, content is poorly or wrongly presented, delivery style is inappropriate (too fast, too slow, unintelligible, etc.)
Two or more evaluators would watch the video and assign it a numerical score in accordance with the rubric in Table 4a.	

Norming

While it is helpful to state criteria clearly, evaluators often vary in how they interpret the criteria. A norming session, in which evaluators review several examples and discuss how they will apply the criteria, will greatly increase reliability as well as helping make explicit what colleagues consider excellent and less than excellent. Criteria are likely to evolve and be refined during the norming process, which is a great opportunity for clarifying values about teaching.

Integrating results

A system that takes into account multiple sources of data and the multiple dimensions of teaching must finally translate the values assigned to each data source into a common language. This allows the systematic synthesis and integration of quantitative and qualitative data.

Again, the process is similar to computing final course grades. A final grade for a course may sum up performance on essay questions, projects, case presentations, multiple-choice exams, and class participation. These very different measures of student competence are integrated by assigning comparable numerical scores to each. Most often, the measures are assigned different weights: for example, a final exam may count for 30% of the final grade while a midterm counts for 15%. The regularized, weighted scores for each component of the final grade are tallied to arrive at a final score.

Integrating the various components of a teaching evaluation is a similar process except that instead of a single grader (the instructor), scoring is typically done individually by a team of evaluators and these results are thenaveraged to produce a summative judgment. This increases reliability.

Each evaluator rates each document, assigning it a numerical score based upon stated criteria and decision rules, as in Tables 4a,b. These are multiplied by pre-assigned weights, following decision rules like that stated in Table 3 for presentation/delivery, resulting in a summary score for each dimension. These are then weighted as previously described and the weights added to arrive at a final score. Table 5 provides an example of this final tallying.

h			
Dimensions	Rating	Weight	Score
Presentation/Delivery	4.6	50%	4.6 x .5 = 2.30
Course design	4.2	30%	4.2 x .3 = 1.26
Effective course evaluation systems	3.5	10%	3.5 x .1 = .35
Mentoring of students	4.7	10%	4.7 x .1 = .47
			Total: 4.38
A minimum of three evaluators review the documents submitted by each candidate and rate them according to stated criteria. Evaluators' scores can be averaged for each major dimension of teaching, or their final summary scores can be averaged to arrive at a final score for each candidate.			

It's too complicated, do we really have to do all this?

You're probably thinking that this all sounds very complicated; why can't we just use student ratings and be done with it? In truth, in most cases, appropriate use of student ratings data will result in the right decisions. However, for the minority of cases where the ratings data may be misleading or incomplete, it is critical to have other measures. In any case, a judgment is more likely to be correct if corroborated by several data sources. Indeed, few faculty members want their teaching evaluation to be based solely on student ratings.

Our university mission, in part based on our land-grant status, states clearly that teaching is important. A department that wishes to promote effective teaching will use evaluation as an opportunity for faculty development. Meaningful and collegial peer evaluation can provide a context for valuing and supporting good teaching.

An imprecise system leaves a department open to litigation, in the worst-case scenario. Short of that, it can have negative effects on department morale. The better the evaluation system, the less likely that faculty will feel they are being judged unfairly or that expectations are unclear. At the same time, a department will be on solid ground in addressing situations of genuine poor teaching.

Chapter 6 Supporting Summative and Formative Evaluation at the Department Level

At best, evaluation provides both an impetus and a reward for developing teaching skills. This is especially true when there are incentives for getting involved and rewards for succeeding. A unit head who promotes the view that evaluation and faculty development are important for everyone, from the best teachers on down, can have an enormous positive effect on teaching in his/her department. The underpinnings of support for effective teaching at the department level are 1) discussion of teaching with new faculty, 2) incentives/rewards for teaching effort and effectiveness, and 3) an inclusive, collegial process for developing and maintaining the evaluation system.

A number of ideas for promoting a positive climate for evaluation and teaching improvement are listed below. To reward participation in these activities, provide "credit" simply for participating.[4] Encourage productive connections between formative activities and summative evaluation by allowing documentation based on formative activities to be used as evidence for summative decisions.

- **Brown Bags with Teaching Themes**

  The simplest method for fomenting thinking about teaching is a regular brown bag lunch series with teaching-related themes. Presenters/facilitators can be department members, invited presenters from other departments, and specialists from the various support services, such as AER, the University Teaching Center (UTC), CCIT, the Library, the University Learning Center (ULC), etc. Sessions can be more or less interactive, potentially involving sharing testing approaches, evaluating tests and assignments, developing learning activities, reviewing syllabi, and making arrangements for reciprocal observation. Faculty who present at these sessions should receive credit, especially if their presentation can be shown to have positive results.

- **Teaching Circles**

  Teaching Circles are small groups of faculty committed to developing their teaching together. They meet regularly and set their own agenda. Many formats are possible; the American Association for Higher Education offers guidelines and activities (Hutchings (1996). Reciprocal peer observation is an activity typically engaged in by teaching circle members (see Bernstein, 1996). For more information, call UTC at 621-7788.

- **Peer Mentoring**

  Faculty, especially junior faculty, are paired with experienced faculty who act as role models and sounding boards for ideas about teaching. Reciprocal peer observation is often part of this arrangement.

---

[4] Bernstein (1996) describes a department in which simple participation in a teaching circle is rewarded. At the same time, observation reports collected over time within the teaching circle turn out to be appropriate documentation of improvement.

- **Master Teachers**

  A large department might consider creating a Master Teacher designation, either as a permanent position or a rotated one. Master Teachers would not teach more, but would become experts on teaching in the discipline. Their responsibilities might involve sharing information with colleagues on teaching-related issues and organizing faculty development activities within the unit. A program of this type currently exists within the Arizona Health Sciences Center. For information about prototypes, contact UTC at 621-7788.

- **Workshops on teaching and evaluation-related topics**

  The various support units offer workshops of different lengths tailored to department needs. These can greatly stimulate a department's thinking about evaluation and teaching, especially if followed up by discussion during brown bags and in teaching circles. AER offers workshops and presentations on many aspects of evaluating teaching and evaluating student learning. Call 626-4214 for more information.

Resources

Bernstein, Daniel J. "A Departmental System for Balancing the Development and Evaluation of College Teaching: A Commentary on Cavanagh. In *Innovative Higher Education*, Volume 20, No. 4, Summer, 1996.

Hutchings, Pat. *Making Teaching Community Property: A Menu for Peer Collaboration and Peer Review*. AAHE, 1996.

Chapter 7 Evaluating Teaching at the Department Level

When teaching is evaluated, the focus is normally the individual instructor. However, a number of trends are making it important to also consider teaching at the unit level. At this level, evaluating teaching should be considered in conjunction with evaluating the curriculum, a broader inquiry with a somewhat different focus. A department could have excellent teachers, yet have a number of instructional problems related to curricular issues: repetition of some material in more courses than necessary, lack of coverage of other important material, courses with ill-defined audiences, etc. Indeed, in many cases, instructional problems are the result of curricular problems rather than deficiencies in teaching skills. The best teacher in the world will find it hard to be fully successful, for example, if a class has two student constituencies with conflicting needs (e.g. majors vs. students taking a course as an elective).

In considering teaching at the department level, summative and formative purposes should be clarified. Summative concerns relate to academic program reviews and other institutional processes. Departments should review their evaluation practice keeping in mind the demands of these institutional reviews. Formative purposes might include optimizing course assignments, deciding whether specialized training for faculty is appropriate, deciding whether courses should be taken consecutively or concurrently, considering possibilities for modularized courses or increased team teaching, etc.

When looking at teaching at the department level, useful areas to focus on are:

- student learning outcomes
- student satisfaction
- faculty satisfaction
- alumni/employer satisfaction
- instructor practices,
- instructor activities related to teaching

AER offers a variety of relevant instruments and reports, detailed below. For assistance with curriculum evaluation or evaluating teaching at the department level, contact aer@email.arizona.edu .

Student Learning Outcomes

While it is difficult to use student learning outcomes in evaluating individual faculty, they are a critical component of the evaluation of teaching at the department level. In fact, a revolution in thinking about the use of student learning data in program evaluation is currently underway. In addition to standardized tests, a wide array of more qualitative and "authentic" approaches are being tried, ranging from performance observations to interview approaches that draw on such techniques as the ethnographic interview. A particularly promising approach is the use of "embedded" instruments, treated as normal parts of individual coursework by students, but examined with a view to understanding programmatic issues.

As with evaluating teaching, the discussion should begin with a broad look at desired outcomes for the major, minor, and casual student. Thought should also be given to assessing students' incoming skills as a prerequisite to determining what they've learned from their programs of study.

A method compatible with some disciplines is using a capstone course as a setting for students to evaluate overall learning in the major. Often, capstone students compile portfolios that include work done for a variety of courses. (In other cases, student portfolios are compiled as students progress through the major (Rogers and Williams, 1999).) Use of portfolios in evaluating department teaching emphases is described in Banta et al (1996) and elsewhere.

Studying samples of student work on designated assignments representing goals for the major can provide an excellent picture of how courses impact students. Results often have implications for the evaluation of individual faculty. Suppose an analysis of student outcomes shows that majors are deficient in certain software skills that potential employers consider desirable. Credit could be offered faculty for designing assignments and projects in which the desired software skills are taught and practiced. Julian (in Banta, 1996) describes a case in which a department, based on review of student portfolios, found its graduates lacking in oral communication skills and used this feedback to stimulate a number of changes in instructional approach.

Student Satisfaction

AER provides a number of reports intended to facilitate an overview of department results. Most importantly, the Comparison Group Summary Reports offer an overview of collective results for each course category within a department. See "Guide to Student Ratings at the University of Arizona" at http://aer.arizona.edu/AER/teaching/Guide/TCEGuide.pdf for more information.

Faculty Satisfaction

Surveys of faculty satisfaction with teaching conditions are unusual, but an excellent source of information that can raise morale and lead to obvious improvements. Understanding what teaching conditions are important to faculty can form the basis for an effective system of incentives. Contact eberman@u.arizona.edu for ideas in this regard.

Alumni/Employer Satisfaction

Many departments have developed survey instruments for soliciting alumni and/or employer opinion about their programs. In general, alumni can provide useful information about a program as a whole and which parts of it have been most and least useful. Since alumni typically represent students' future paths, they are a good source of data about whether students will find their preparation adequate. Employers can provide information about whether they are satisfied with the skills and knowledge levels of recent graduates.

Instructional Practices and Priorities

The Course Profile is a survey instrument designed to collect information about instructional goals, instructional activities, and aspects of how students are evaluated. (It may be seen at [web address]). Using the Course Profile within a department will provide information about a number of aspects of teaching practice. Call 621-9585 for more information.

Teaching-Related Activities

A department might want to evaluate the extent to which resources for improving teaching are used and how well they are working. Tracking participation in department-, college-, and university-sponsored activities provides data for making decisions about resources as well as potentially impacting criteria for judging individual faculty. For example, if faculty participating in a department-based teaching circle program are receiving, on average, higher student ratings than faculty not participating, a decision might be made to increase "credit" for participation as an incentive.

Sources and References

Anderson, R.S. and B.W. Speck. *Changing the Way We Grade Student Performance: Classroom Assessment and the New Learning Paradigm*. San Francisco: Jossey-Bass Publishers. New Directions for Teaching and Learning, No. 74, Summer 1998.

Angelo, T.A. and K. P. Cross. *Classroom Assessment Techniques.* Second Edition, San Francisco: Jossey-Bass Publishers, 1993.

Arreola, Raoul A. *Developing a Comprehensive Faculty Evaluation System*. Bolton, MA: Anker Publishing Company, Inc., 1995.

Banta, Trudy et al. *Assessment in Practice.* San Francisco, Jossey- Bass Publishers, 1996.

Bernstein, Daniel J. "A Departmental System for Balancing the Development and Evaluation of College Teaching: A Commentary on Cavanagh." In *Innovative Higher Education*, Volume 20, No. 4, Summer, 1996.

Braskamp, Larry A., and Ory, John C. *Assessing Faculty Work*. San Francisco, CA: Jossey-Bass Publishers, 1994.

Brinko, K. T. and R. J. Menges, eds. *Practically Speaking: A Sourcebook for Instructional Consultants in Higher Education*. Stillwater, Okla: New Forums Press, Inc., 1997.

Centra, John A. *Reflective Faculty Evaluation*. San Francisco, CA: Jossey-Bass Publishers, 1993.

Cohen, P.A. "Student Ratings of Instruction and Student Achievement: A Meta-Analysis of Multisection Validity Studies." *Review of Educational Research*, 51, pp. 281-309, 1981.

DeZure, Deborah, "Evaluating Teaching through Peer Classroom Observation," in Peter Seldin and Associates, *Changing Practices in Evaluating Teaching*, Bolton, Mass.: Anker Publishing Company, 1999.

Diamond, Robert M. *Preparing for Promotion and Tenure Review*. Bolton, MA: Anker Publishing Company, Inc., 1995.

Diamond, Robert. *Designing and Assessing Courses and Curricula in Higher Education*. San Francisco: Jossey-Bass Publishers, 1998.

Doyle, Kenneth O. *Evaluating Teaching*. Lexington, MA: D.D. Heath and Company, 1984.

Hutchings, Pat, Ed. *From Idea to Prototype: The Peer Review of Teaching*. Washington, DC: American Association for Higher Education, 1995.

Hutchings, Pat. *Making Teaching Community Property: A Menu for Peer Collaboration and Peer Review*. AAHE, 1996.

Jacobs, L.C. and C.I.Chase. *Developing and Using Tests Effectively: A Guide for Faculty*. San Francisco: Jossey-Bass Publishers, 1992.

Keig, Larry and Michael D. Waggoner. *Collaborative Peer Review: The Role of Faculty in Improving College Teaching*. ASHE-ERIC Higher Education Report No. 2. Washington, DC: The George Washington University, School of Education and Human Development, 1994.

Murray, Harry G. "Effective Teaching Behaviors in the College Classroom." In J. Smart (Ed.), *Higher Education: Handbook of theory and research* (Vol 7, pp. 135-172), New York: Agathon Press, 1991.

Murray, John P. *Successful Faculty Development and Evaluation: The Complete Teaching Portfolio.* ASHE-ERIC Higher Education Report No. 8. Washington, D.C.: The George Washington University, Graduate School of Education and Human Development. 1995.

Ory, J.C. and K.E. Ryan. *Tips for Improving Testing and Grading.* Newbury Park, CA: Sage Publications, Inc., 1993.

Rogers, G.M. and J. Williams. "Building a Better Portfolio. " In *ASEE Prism*, January 1999.

Seldin, Peter. *The Teaching Portfolio*. Bolton, MA: Anker Publishing Company, Inc., 1997.

Walvoord, B.E. and V.J. Anderson. *Effective Grading: A Tool for Learning and Assessment.* San Francisco: Jossey-Bass Publishers, 1998.

Weimer, Maryellen, J.D. Parrett, and M. Kerns. *How am I Teaching: Forms and Activities for Acquiring Instructional Input.* Madison, WI: Magna Publications, Inc., 1988.

Web Sources

An Introduction to Classroom Assessment Techniques:
http://www.psu.edu/idp_celt/CATs.html

Classroom Assessment Technique Examples:
http://www.hcc.hawaii.edu/intranet/committees/FacDevCom/guidebk/teachtip/assess-2.htm

Classroom Assessment Techniques:
http://www.hcc.hawaii.edu/intranet/committees/FacDevCom/guidebk/teachtip/assess-1.htm

Classroom Assessment Techniques: http://www.ntlf.com/html/lib/bib/assess.htm

Classroom Assessment Techniques in the Sciences: http://www.flaguide.org/

Guide to Student Ratings at the University of Arizona:
http://aer.arizona.edu/AER/teaching/Guide/TCEGuide.pdf

Peer Observation of Teaching: http://www.ltsn.ac.uk/genericcentre/index.asp?id=17849

Appendix A Evaluating TCE Results[5]

Academic units should have a written policy detailing how TCE results are evaluated for purposes of administrative review. AER recommends a three-part process consisting of 1) evaluating the sample, 2) reviewing results and assigning points according to a rubric, and 3) verifying the results of the review by examining the candidate's narrative and taking into account mitigating factors. Guidelines for reviewing the sample, a generic rubric for examining results, and a sample statement of ratings adjustments are provided below.

1. Evaluate the Sample

Ratings results should be used in summative evaluation only if they are representative. The higher the proportion of respondents to those enrolled, the more reliable the results. In general, sections with a less than 50% response rate should not be used for performance appraisal. The smaller the class, the higher the percentage of responses needed to ensure that the same is representative.

One way to ensure reliability is to assign each section a "sample score" based on the percentage responding, then average the scores for each level of course (lower division, upper division, graduate) to arrive at a sample score. Samples not meeting a specified level should not be considered in summative review. Table 1 below provides suggested sample scores for different enrollment sizes, while Table 2 offers interpretations for averaged sample scorees.

Table 1. SUGGESTED TCE "SAMPLE SCORES"

0=poor sample
1=marginal, but likely usable
2=probably good sample

Enrolled	Response %	Section Sample Score
5-29	Less than 50%*	0
	More than 49%, but less than 80%	1
	More than 79%	2
30-49	Less than 50%*	0
	More than 49%, but less than 75%	1
	More than 74%	2
50 or more	Less than 50%*	0
	More than 49%, but less than 66%	1
	More than 66%	2

* These results are considered unusable because it cannot be determined if the few students who responded were representative of the class as a whole.

[5] An earlier version of this Appendix, entitled "Preparing a Quantitative Summary of TCE Results," was co-written with Jennifer Franklin.

Table 2. Mean Grad and Undergrad Sample Scores	
Values	**Interpretation**
2.0 to1.5	Good sample across all sections
1.49 to.50	Marginal, but likely usable
.50 to 0	Unusable set of sections; too few respondents for reliable interpretation

Inadequate sample scores may be addressed in the narratives faculty write to accompany the "quantitative summaries" they are expected to provide for administrative reviews. AER recommends that departments exclude from further consideration ratings results where the sample is inadequate (Section Sample Score equals 0; Summary Sample Score is less than .50).

Part 2. Evaluate TCE Results

Department plans for faculty performance appraisal should include an explicit (written) statement of the basis for judging TCE results. Essentially, there are two choices: criterion-based or norm-based. In criterion-based schemes, the performance of individuals is compared with fixed standards (e.g. ratings over 4.5 are deemed "outstanding"). In a strong teaching department, everyone could be deemed outstanding or excellent since individual scores are not affected by the scores of others. In norm-based schemes, the performance of individuals is compared with that of their peers (e.g. the top 10% of ratings are deemed "outstanding"). Norm-based schemes are conceptually similar to grading on the curve in that standards are relative to that of peers rather than absolute.

After determining whether a norm-based and a criterion-based approach is chosen, explicit "decision rules" for interpreting ratings should be developed, as in Table 3 below. Ideally, decision rules should be a matter of department policy. They can include guidelines for incrementing scores under certain conditions (see below).

Table 3. SECTION TCE SCORING CRITERIA (criterion-based*)		
Suggested Criteria:	**Finding**	**TCE Points**
Most ratings** between 4.5 and 5.0	Exceeds unit criterion (outstanding)	5
Most ratings between 4.0 and 4.5	Meets or exceeds unit criterion (excellent)	4
Most ratings between 3.5 and 4.0	Meets unit criterion (good)	3
Most ratings between 3.0 and 4.0 of scale	Meets unit criterion , but some improvement is desirable (needs improvement)	2
Most ratings below 3.0	Does not meet unit criterion and substantial improvement is required (unacceptable)	1
Ratings problematical due to high CIs, insufficient participation, etc.		***

* In some departments, norm-based systems are inappropriate because there is too little difference between the bottom and the top. In general, norm-based systems work best when there is a wide range of variation in results.

**not including text/readings and course difficulty items

***these may be either excluded or decided on by the group of evaluators

Part 3. Adjust Results

Relying only on decision rules may lead to unfair judgments. For example, a large required upper division course may receive relatively low ratings compared to ALL upper division courses, but normal or even high ratings compared to other LARGE upper division courses. Because size is not taken into account in the comparison groups used in Overall Effectiveness Graphics, a person teaching large courses could be at a disadvantage if numbers alone are considered. Low ratings may also occur because an instructor is experimenting with a new approach and runs into unexpected problems, or due to factors the instructor cannot control. Faculty should detail special circumstances in a narrative that accompanies their presentation of quantitative results.

AER recommends that units explicitly describe how they will treat special circumstances. Ratings can be adjusted by assigning "bonus values" or increments. Table 4 offers an example of a statement of rating adjustments.

Table 4. Sample Ratings Adjustments

Circumstance	Increment*
New Course Increment: for courses being taught for the first time	+.5
Innovation Increment: for courses in which new instructional methods valued by the unit or college are being introduced	+.5
Challenge Increment: for classes rated significantly higher in difficulty than the comparison group **and** which have high ratings (This provides incentive for **not** inflating grades.)	+.25
Special Circumstances Increment: for courses where circumstances beyond an instructor's control led to lower ratings than would have been otherwise merited (based on instructor's usual ratings), e.g.	
• inadequate instructional facilities or resources	+ .2
• an unusually large number of unprepared or poorly qualified students were enrolled in the course	+ .2
• a personal circumstance in the instructor's life (e.g. illness or a death in the family)	+ .5

Appendix B: Using Student Written Comments in Summative Evaluation[6]

UA P&T policy guidelines list a "summary of student interviews or comments on questionnaires" as a required part of documenting teaching effectiveness. Because there are many potential difficulties in using student comments fairly, evaluation experts recommend precautions due to validity, reliability, and generalizability concerns. Below are some suggestions for preparing this part of the tenure or promotion dossier.

Student comments are typically drawn from responses to end-of-semester student ratings questionnaires (TCEs). However, comments written on TCE questionnaires are problematical because 1) usually only a small percentage of respondents provide them, 2) these respondents often have stronger-than-usual opinions (both positive and negative), and 3) comments typically refer to any and all aspects of the course, so there is unlikely to be consensus on any specific aspect.

In summative reviews such as promotion and tenure decisions, comments should reflect a reasonably representative sample of student opinion in terms of both content and quantity. If comments are provided by less than half the class, it is preferrable to use other methods for collecting student comments, such as instructor-designed surveys (administered when all or most students are present) and Midsemester Focus Group Evaluations.

Minimum requirements for a fair summary include: 1) an indication of the number and percentage of students writing comments compared to those who submitted TCE questionnaires, 2) a discussion of the issues addressed and the number of positive and negative comments on each issue, and 3) a few examples of representative positive and negative comments. The summary should fairly represent all comments.

Summarizing raw written comments requires interpretation, paraphrasing, and reduction. Single written comments and even patterns of comments may be understood differently by different interpreters and evaluators may be unduly influenced by a single highly articulate opinion, positive or negative. If comments are reviewed by a committee, they should be transcribed to avoid handwriting recognition, and an approach to analysis agreed upon. Using multiple readers and systematic content analysis methods will go a long way toward ensuring validity and reliability.

The bottom line is that an academic unit's case would be weakened if an instructor litigated an adverse personnel decision (such as denial of tenure or an allegedly inequitable merit raise) and showed that the decision-makers had been improperly influenced by student written comments.

Suggestions for Using Student Written Comments in Summative Evaluation

- have an explicit policy concerning the use of written comments and make sure it is applied fairly and consistently
- protect student anonymity by transcribing comments before they are reviewed and deleting any references potentially attributable to individual students
- collect a reasonably representative sample of student opinion from a reasonably representative sample of the instructor's teaching load during the period under review

---

[6] An earlier version of this appendix, co-written with Jennifer Franklin, was published in ***Instructional Evaluation and Faculty Development***, 18(1), 1998.

- analyze and summarize the comments using a valid and reliable content analysis strategy (i.e. multiple readers, "norming" procedures for interpretation, a reasonable estimate of inter-rater reliability)
- use the summaries of the comments, not the "raw" comments, for evaluation purposes
- take precautions to prevent the introduction of anecdotal information based on comments seen in formative evaluations or heard in passing

Appendix C: Possible Items for the Teaching Portfolio

Documenting course design and development

- course portfolio (a collection of materials demonstrating effort and results relating to a single course; could include all the bulleted items, as well as examples of student work.)
- course plan (a thematic outline of a course connected to an explanation of how the various course components and activities fulfill the objectives of the course. The course plan should cover: selection/organization of content, instructional strategies (in-class activities, assignments, course materials), evaluation and analysis of student outcomes and satisfaction.)
- syllabus
- website
- study guides and other course materials
- courseware
- evidence of activities instrumental to course design including list of books/materials reviewed, colleagues consulted, workshops attended
- exams and assignments
- examples and analyses of exams and assignments
- evidence of formative evaluation

Documenting course delivery

- student ratings results
- peer or expert observation reports
- video of classroom presentation
- teaching awards or nominations

Documenting course administration and maintenance

- Peer evaluation
- Self-provided data sheet

Documenting faculty development/teaching scholarship

- grants/awards for pedagogical activities
- articles on pedagogy
- workshops/conferences attended
- workshops/conferences presented at
- participation in department/college/university/disciplinary faculty development programs
- documentation of results of test analysis, classroom research

Documenting advising and mentoring activities

- student ratings of advising
- self-provided data sheet

Appendix D: Using Peer Observation in Summative Evaluation

While peer observation is an excellent tool for formative evaluation (see p. 5), for summative evaluation, it must be embarked upon with great care. Research over many years, indeed decades, consistently shows peer observation to be far less reliable and far more subject to invalidating factors than student ratings of instruction.[7] And this is not surprising: "peer observation" frequently means a single visit by an individual untrained in observation or instructional evaluation, without even a protocol for guidance. Judgments frequently reflect preconceptions about effective instruction and/or about the individuals being observed. They often focus on content coverage, with little relation to instructional effectiveness vis-à-vis the students, a validity problem. Results often lack reliability because there is no basis for knowing whether the observed class is characteristic or unusual. Evaluation procedures with validity and reliability problems, when used as the basis for personnel decision-making, open the door to lawsuits as well as bad feelings within a unit. While peer observation for formative purposes typically enhances collegiality, peer observation for summative purposes can be extremely destructive of collegiality.

A valid and reliable system of peer observation for summative purposes is not impossible. What it requires is time and commitment. A minimum of three observers each observing three classes would likely ensure reliability. (One way to accomplish this is to videotape three classes and have the observers review the videotapes. This ensures that evaluative judgments are in response to the same source material.) To increase interrater reliability, a common evaluation protocol should be used, and evaluators should participate in training and norming sessions. The evaluation protocol should be general enough for individual results to be compared, yet broad enough to encompass the wide range of teaching approaches used within a department. For example, an evaluation protocol designed for evaluating lecture courses would probably not yield valid results if applied to classes based largely on cooperative learning. Without the sort of commitment suggested by these concerns, evaluation experts recommend against using peer observation for summative evaluation.

AER offers training in observation for both formative and summative purposes. A three-hour workshop is available upon request that addresses the benefits and limits of observation, using protocols, and tailoring processes for formative and summative purposes. Participants view several videos of instructors, practice using observation protocols, and share their responses. Departments interested in peer observation of teaching may consult AER at 621-9585 for example processes and protocols. An excellent summary of the issues may be found in DeZure (1999).

---

[7] For example, Doyle (1984) cites a comparison of student and colleague ratings in which "Centra (1975) projected an interrater reliability of .85 for a group of fifteen students, but a reliability of only .57 for a group of fifteen colleagues. "

Appendix E: Developing an Academic Unit Program for Evaluating Teaching: A Checklist[8]

This list appears linear, but many of the activities occur simultaneously, with much back-and-forth among them.

I. Decide on a procedure

Who will develop the program? An individual? A committee? How will input be solicited? Who will say yes or no to the proposed program? What are the timelines for development and implementation? What resources will be allocated to ensure that the program is implemented?

II. Design the program

Review the existing procedure

How adequate is it? What does it fail to cover? Should it be scrapped or modified? How does it fit into the overall faculty evaluation program?

Define policies for administering the program

How will working committees be constituted and charged? Who will maintain files? Who will review evaluation packets? What will the timetable and procedure be for faculty to submit documentation and receive a response? How will unit policy relate to college and university policy?

Define what will be evaluated

What roles, aspects, and activities of teaching and teaching related functions will be evaluated?

Set criteria for each aspect of teaching being evaluated

How will you know what is excellent, satisfactory, less than satisfactory? Evaluation is fundamentally a process of deciding what is valued. Criteria of merit must be specified for each aspect of teaching that will enable informed observers to make consistent judgments about excellent, satisfactory, and deficient performance. Worth, or instrumental value (e.g. number and types of courses taught), should also be taken into account.

Set weights for each aspect of teaching being evaluated

How will the various aspects of teaching be weighted? How will especially valued activities (i.e. those central to the unit's mission) be recognized? A weighting schema for different aspects of teaching provides flexibility and a fair recognition of the wide range of instruction-related activities faculty actually engage in. It also makes it possible to reward both merit and worth.

Identify valid sources of data for each aspect of teaching being evaluated

What (who) are the best (i.e. most valid, most reliable, most useful, most practical) sources of data for each aspect of teaching? Because instruments for evaluating teaching are subject to error, the more independent sources of data, the more reliable the program will be.

---

[8] This Appendix was co-written with Jennifer Franklin.

Develop procedures for collecting data

How will data collection procedures be standardized? What will be used as evaluation instruments? How will issues of validity and reliability, essential to ensuring a fair and legally defensible evaluation system, be addressed? Who will do what when?

Develop protocols for interpreting data and making decisions

How will data be translated into decisions and actions? That is, what decision rules will be used to interpret quantitative and qualitative data and how will interpretations be applied in decision-making processes?

Develop a training program for evaluators

How will faculty learn to use evaluation data and make decisions? What will ensure that evaluation practice is consistent from year to year and that evaluators follow common procedures?

Design a process for faculty needing to improve performance and allocate resources

Continuing Review documents require that a process be designed for addressing deficiencies in teaching.

Develop an appeal process

Design a plan for evaluating the system

How will the unit know that faculty and administrators are following the plan in a fair and consistent manner? How will problems be identified?

III. Pilot-test the program and make revisions as needed.

IV. Evaluate the program on a regular basis

In the early stages, reviewing the program annually may be helpful. Once established, the program should be evaluated every few years. The feedback from the evaluation process may show when additional support such as training of faculty evaluators or revisions of key materials is needed.