THE PROFESSIONAL EVALUATION OF TEACHING

James England
Pat Hutchings
Wilbert J. McKeachie

American Council of Learned Societies

ACLS OCCASIONAL PAPER, No. 33
ISSN 1041-536X

THE PROFESSIONAL EVALUATION OF TEACHING

James England
Pat Hutchings
Wilbert J. McKeachie

American Council of Learned Societies

ACLS OCCASIONAL PAPER, No. 33

Introduction

The American Council of Learned Societies is committed to the professional model of the teacher-scholar: the teacher who is devoted to constant exploration, the scholar who places research in the context of challenging the next generation.How the one activity serves the other was explored by Francis Oakley in ACLS Occasional Paper No. 32, *Scholarship and Teaching: A Matter of Mutual Support.*

Within the community of scholars we have developed widely accepted procedures, grounded in peer review, for evaluating scholarship. We use these procedures to improve works while they are still in manuscript, to make judgments about what to publish, and to provide a basis for personnel decisions. There is less agreement on how we should evaluate teaching for purposes of improvement or judgment. What are appropriate professional approaches to the evaluation of teaching?

This Occasional Paper on *The Professional Evaluation of Teaching* originated in presentations at the 1996 Annual Meeting in Washington, D.C. Wilbert J. McKeachie, professor of psychology at the University of Michigan, and delegate to the ACLS from the American Psychological Association, summarizes a considerable body of research on student evaluations, including work of his own, showing that student evaluations have considerable validity and are not subject to a number of biases of which they are frequently suspected.

Pat Hutchings, director of the Teaching Initiatives Group of the American Association for Higher Education (AAHE), reports on a project she directs on peer review of teaching. She stresses a primary focus on establishing a culture that nurtures the improvement of teaching through peer collaboration. A dozen institutions are participating in the pilot project, as are several learned societies which belong to the ACLS.

James England, provost of Temple University and former provost of Swarthmore College, addresses how evaluations of teaching are used in personnel decisions. He supports using both student- and peer-evaluation approaches (Temple University is participating in the AAHE Project); McKeachie and Hutchings also see student and peer evaluation as complementary, not competing, approaches. In addition, England sketches a third approach, one that would focus on assessments of what students actually learn.

We hope this Occasional Paper will provide a useful perspective on what constitutes appropriate professional evaluation of teaching for faculty, deans, department chairs, board members, and others.

Contents

Student Ratings of Teaching

Wilbert J. McKeachie
University of Michigan

In 1946, when I began teaching at the University of Michigan, the faculty had already voted that student ratings of teaching should be collected in all classes, and I used student ratings along with data on achievement in my first research study on effective teaching in the winter of 1947.

In 1949 Dean Heyward Keniston gave me a graduate assistant to collect data and review research on student ratings. We found that student ratings were collected at a number of universities and colleges, including Harvard, the University of Washington, and Purdue.

In 1951 the program for evaluating teaching came up for review. I remember the heated debate in the College of Literature, Science and the Arts about the recommendation of the College Executive Committee that the college continue to require collection of student ratings in all courses. Some faculty members felt strongly about the impropriety of students presuming to express opinions about a professor's teaching. Encouraging students to think that they were qualified to make judgments about teaching would destroy the proper respect students should have for the faculty. Others asserted with great fervor that teaching is an art; it is impossible to evaluate in terms of some form of measurement.

The result of the debate was the adoption of an open-ended form to be administered in all courses. There were five questions:

1. What do you think are the objectives of the course?
2. What is your judgment of the value of this course in your education? Please point out both its contributions and its deficiencies.
3. To what extent did you learn to think critically in the area covered by this course?
4. Assuming you have the time and opportunity either in college or later, do you think you would be inclined to pursue interests in this area? Explain.
5. Keeping in mind that the returns from this questionnaire will be used by the instructor in improving his teaching, please mention any other aspects of the course or instructor (such as, for example, clarity of presentation) not covered in previous questions which you consider to be especially good or poor, and offer any suggestions you have for the improvement of the course.

As I look at this almost half a century later, I'm impressed. I don't think we do any better today.

During the period of student activism in the 1960s and 1970s, there was a great increase in the use of student ratings. As Pat Hutchings indicates, they are now used in most colleges and universities.

Nonetheless student ratings are still controversial. Most of us are sensitive about being evaluated, and anytime the results are negative it is natural to question the validity of the evaluation. And there are some negative evaluations in almost every class. That's not surprising; teaching that is effective for some students is not equally effective for everyone.

Some years ago I was a member of a committee administering grants to senior faculty members who proposed to construct or modify their courses to emphasize thinking. At an end-of-the-year dinner for the participants, the discussion turned to student ratings, and the usual criticisms were raised.

"Students don't really appreciate a good course until they are out of college."

"Students can't really judge how well they are learning."

"Students only give high ratings to courses with low standards."

It happened that Herb Marsh, a professor at the University of Western Sydney, was visiting me at the time, and I had invited him to be my guest at the dinner. He is probably the world's leading researcher on student ratings of teaching, and as a guest he kept quiet as long as he could. But finally he could stand it no longer and said, "You know, there's a good deal of research evidence on the issues you've raised."

A prominent historian immediately retorted, "We don't care about research evidence; we have our own experience."

So much for teaching critical thinking!

In any case, I have heard Kenneth Feldman (the preeminent reviewer of research on student ratings) say that there have been over 2,000 articles published on student ratings—well over 1,000 of which present research evidence. In fact, we probably have more good research on student ratings than on any other aspect of higher education.

There are three major uses of student ratings:

1. Student guidance in choice of courses.
2. Improvement of teaching.
3. Evaluating teaching for use in personnel decisions; e.g., tenure or merit salary increases.

How well do student ratings achieve these purposes?

Student Guidance

This is first by right of seniority. Student ratings were first collected, I believe, at Harvard University and published to provide guidance for students in choosing courses. Faculty members usually presume that students are thus likely to choose the easiest courses, but in a study we ran several years ago, we found that, as compared with an uninformed control group, students given student ratings of two alternative introductory courses chose the more highly rated course, even though it was rated as requiring more work (Coleman and McKeachie).

Improvement of Teaching

Harvard was not alone in using student ratings. In the mid-1920s Herman Remmers of Purdue University began a program of research on student ratings that made substantial contributions for over four decades. His studies are still among the best that have ever been done.

Remmers and his students found:

a. In multi-section courses, the teachers of those sections achieving higher scores on classroom examinations are rated higher than those teachers whose students have not learned as much. Moreover, if a teacher aims a course at the top students, those students give higher ratings than the rest of the class. However, if a teacher is particularly effective with the poorer students those students rate the teacher higher (Elliott).
b. Ratings of teachers by alumni 10 years after graduation correlate well with ratings of the same teachers at the end of a course (Drucker and Remmers).
c. Student characteristics such as age, sex, class standing, and grade in the course have little effect on ratings of teaching (Remmers and Brandenburg).

The research of Remmers and those who have followed him also strongly indicates that:

d. Student ratings returned to faculty members result in some improvement of teaching, but not very much.
e. There is more improvement if behavioral items are used rather than more abstract, general terms. For example, instead of asking about clarity, ask "Uses concrete examples" or "Fails to define new terms" (negative), or instead of an item on organization, use "Reviews topics from previous lecture" or "Puts outline on the blackboard" (Murray).

f. There is substantial improvement when the ratings are discussed with another teacher (McKeachie et al. 1980).

Personnel Decisions

If student ratings are part of the data used in personnel decisions, one must have convincing evidence that they add valid evidence of teaching effectiveness. I have already reviewed Remmers's extensive validity studies. They have been replicated at other universities. In general, better teachers (as measured by student learning) are rated higher by students. In addition, there is evidence that students of faculty members who are rated highly are more likely to be motivated to further learning as indicated by election of advanced courses in the same field. Highly rated teachers also produce more change in attitude sophistication (McKeachie, Lin, and Mann). The instructor's own judgment also correlates well with student ratings. Marsh found that if you asked instructors which of two classes went better, their judgments agreed well with the student ratings of the classes.

Finally, for this assembly, perhaps the most interesting evidence of validity is that humanities teachers are rated as being more effective than teachers of science, math, and engineering (Feldman). There is also fairly persuasive evidence that humanities teachers are actually better teachers. Humanities teachers:

1. Are more expressive—move around more, use more gestures.
2. Know students' names, encourage questions, ask questions.
3. Show an interest in student ideas, show concern for student progress.
4. Ask more questions requiring analysis and synthesis on exams; science and technology teachers ask more rote memory questions (Murray & Renaud).

These are characteristics that lead to longer-term retention and greater gains in thinking and motivation.

But aren't there biases or contextual factors that can invalidate student ratings? Probably the most common criticism by uninformed faculty members is that you get good ratings by "dumbing down" your course. Cutting down the amount of work will, they feel, inevitably result in higher student ratings. The facts, however, indicate that this is not generally true. Student ratings of teaching are higher for courses that are rated as requiring more work or that are more difficult. Undoubtedly, there is a limit. If a course is pitched above the students' heads, or if the

course requires more work that most students can do, so that less learning results, student ratings will be lower than for courses that result in better learning.

Generally, small classes are rated higher than large classes, but research shows that small classes are more effective than large classes in producing changes in thinking, motivation, and attitudes (McKeachie). Similarly, there are often small differences between required and elective classes and lower-level vs. higher-level classes.

But the great concern about bias is based on the idea that we should be able to compare teachers to one another, that the number 3.1 should signify better teaching than 3.0, that we should be able to compare two teachers teaching in different departments at different class levels with different students. I argue that this is neither necessary nor desirable. In fact, for promotion and salary decisions we do not need to make such comparisons. For the decision about promotion we really only need two categories—good enough to promote or not promotable. Even for salary increases we need only to determine whether the teacher is excellent, good, adequate, or in need of help. We can determine these categories simply by looking at the distribution of the student responses. What proportion of the students give favorable ratings?

We don't need to figure averages to a decimal point. Comparing teachers with averages such as 4.3 and 4.1 is like comparing apples to oranges. We can tell a good apple or a good orange, but judging whether a good apple is better than a good orange is a much more difficult task.

Conclusion

It is clear that student ratings have the potential to contribute positively both to improvement of teaching and to the quality of personnel decisions about teaching. The problem is not in the ratings but in their use.

Students. Student time is used to fill out the ratings, but the students get little benefit from the time they invest. They are not encouraged to think about their own learning and their own responsibility for learning. Answering the questions should be an educational experience, not a mindless appraisal of the teacher.

Forms. The forms used in many colleges and universities are not as useful as they could be. Often a college or department requires that a set of standard items be used. Typically such items are not as applicable to the specific course as would be the case if the teacher developed or chose items specifically about aspects of the course. Moreover, the very

fact that the items are mandated is likely to lead to resentment and resistance.

Norms. In order to conclude that a teacher is reasonably good, what percentage of his or her students would you expect to rate the teacher as excellent? Ten percent? Twenty percent? Fifty percent? Certainly, if at least half of your students think you are excellent, you can't be too bad. At the University of Michigan, over 90 percent of the faculty are rated as excellent by the majority of their students; yet when the faculty members look at their results, almost half of those rated as excellent by the majority of their students find that they are below average. This is discouraging and is more likely to result in a loss of motivation than in increased enthusiasm for teaching.

Evaluators. Whatever the source of data—student ratings, peer evaluation, gossip—some committee or administrator has to make an evaluative judgment. Students are not the evaluators; they simply provide data to the evaluators. In most universities the initial evaluation is made by peers—faculty members elected or appointed to a committee that reviews the evidence for promotion or merit increase in salary.

A key element is the good sense of the evaluators. Unfortunately, many evaluators have stereotypes about what constitutes good teaching, despite the fact that there are many ways to be effective. Thus, they may undervalue a teacher because the students' judgments of their own learning may not fit with the pattern of ratings on such characteristics as organization or enthusiasm, or other characteristics usually associated with effective teaching.

Often the evaluators give less weight to the student ratings than to less dependable evidence, such as peer observations of teaching, testimonials, or general impressions of the teacher's personality.

In an effort to be objective, the evaluators may substitute arbitrary criteria for reasoned judgment. Thus, they may set as a criterion for promotion such that the teacher must exceed a certain numerical mean on student ratings, without consideration of what the teacher is trying to accomplish, the circumstances under which the teacher has to work, the kind of course being taught, the nature of the students, and the many contextual factors that should temper their judgment.

What can we conclude? As Pogo (or one of the "Pogo" characters) said, "The enemy is us."

Coleman, J., and W.J. McKeachie. "Effects of Instructor/Course Evaluations on Student Course Selection." *Journal of Educational Psychology* 73 (1981): 224-26.

Drucker, A.J., and H.H. Remmers. "Do Alumni and Students Differ in Their Attitudes Toward Instructors?" *Journal of Educational Psychology* 42 (1951): 129-43.

Elliott, D.N. "Characteristics and Relationships of Various Criteria of Teachings." Diss. Purdue University, 1949.

Feldman, K.A. "Course Characteristics and College Students' Ratings of Their Teachers: What We Know and What We Don't." *Research in Higher Education* 9 (1978): 199-242.

Marsh, H.W. *Students' Evaluations of University Teachings: Research Findings, Methodological Issues, and Directions for Further Research*. Elmsford, NY: Pergamon, 1987.

McKeachie, W.J. *Teaching-Tips: Strategies, Research, and Theory for College and University Teachers*. 9th ed. Lexington, MA: D.C. Heath, 1994.

McKeachie, W.J., Y-G Lin, M. Daugherty, M.M. Moffett, C. Neigler, J. Nork, M. Walz, and R. Baldwin. "Using Student Ratings and Consultation to Improve Instruction." *British Journal of Educational Psychology* 50 (1980): 168-74.

McKeachie, W.J., Y-G Lin, and W. Mann. "Student Ratings of Teaching Effectiveness: Validity Studies." *American Educational Research Journal* 8 (1971): 435-45.

Murray, H.G. "Low-Inference Classroom Teaching Behaviors and Student Ratings of Teaching Effectiveness." *Journal of Educational Psychology* 75 (1983): 138-49.

Murray, H.G., and R.D. Renaud. "Disciplinary Differences in Classroom Teaching Behaviors." *Disciplinary Differences in Teaching and Learning: Implications for Practice. New Directions in Teaching and Learning*. No. 64. Ed. N. Hativa and M. Marincovich. San Francisco: Jossey-Bass, 1995.

Remmers, H.H., and G.C. Brandenburg. "Experimental Data on the Purdue Rating Scale for Instructors." *Educational Administration and Supervision* 13 (1927): 519-27.

The Peer Collaboration and Review of Teaching

Pat Hutchings
American Association for Higher Education
Teaching Initiatives Group

My piece of this picture, as I understand it, is to talk about the role of faculty in the evaluation of teaching—peer review, if you will. I do so in the context of a national project I've been involved with for the past several years, a project of 12 universities, working in pilot departments, coordinated by the American Association for Higher Education (AAHE), in partnership with Lee Shulman at Stanford University, and funded by the Pew Charitable Trusts and the William and Flora Hewlett Foundation—which pretty much captures all the vital statistics in one sentence. The project, entitled "From Idea to Prototype: The Peer Review of Teaching," was launched in January of 1994 at the AAHE's National Conference on Faculty Roles and Rewards, largely in response to emerging campus recommendations—first in the 1991 "Pister Report" at the University of California, but now widely heard—that teaching, like research, should be peer-reviewed; the intent was to help campuses move toward peer review together, and to ensure faculty involvement, from the outset, in shaping strategies for peer collaboration and review that would be intellectually rigorous, appropriate to the disciplines, and of practical use in improving the quality of teaching and learning.

Context and Rationale

During the 1970s and 1980s, clear progress was made in the evaluation of teaching; student ratings of teacher effectiveness, once the exception, became the rule, and some 86 percent of liberal arts campuses now routinely require that student ratings be used in the evaluation of teaching (Seldin). The next step, it would seem, the next stage of evolution in our seriousness about teaching is to make teaching—like research—a subject for peer collaboration and review.

There are a number of arguments for doing so, of which I'll mention only three (and briefly) here. First, student evaluations of teaching, though essential, are not enough; there are substantive aspects of teaching that only faculty can judge and assist each other with. Currency in the field is an obvious example, the setting of appropriate standards for student work, another. The aim of peer review, let me hasten to say,

is not to replace or supersede evidence provided by students, but to augment and enrich the picture we get from that traditional source. Indeed, many of the strategies being explored in the AAHE project entail ways that faculty peers can help each other gather better and more useful information from students and about learning, for instance, through focus groups, interviews with one another's students, and "co-assessment" of student work.

Second, peer review of teaching is important because teaching entails learning from experience, which is difficult to do without colleagues. It's difficult because to learn from experience, one must have a clear view, and that's hard to get in the booming, buzzing confusion of the classroom. Faculty can help one another step back and see more clearly, and therefore learn from, their own teaching practice in a variety of ways—through direct classroom observation, videotape, and collaborative case studies of teaching practice, to name three possibilities.

Third, and perhaps most important, peer review puts faculty in charge of the quality of their work as teachers. As things now stand on many campuses, the evaluation of teaching at least feels like something that happens *to* faculty: The evaluation forms get delivered to class, filled out by students, and shipped off to the dean's office; or the department chair parachutes into class one day, checklist in hand, to conduct an observation. Indeed, even the method of improvement—in the form of "faculty development"—tends to treat faculty as objects; as a wry faculty friend of mine put it recently (speaking from a campus which shall remain nameless), "We're developing faculty to death." Against this current reality, the argument for peer collaboration and review is that it's a process, or perhaps a set of practices, through which faculty can be active agents—rather than objects—in improving and ensuring the quality of their work at teaching. That is the right and professionally responsible thing to do; it's also a smart one, for if faculty don't oversee the quality of teaching, outside, bureaucratic forms of accountability, already very much in the air, will surely rule the day.

Lessons from the AAHE Project

There are now lots of interesting stories and examples from the work of faculty in the AAHE peer review of teaching project (Gabaccia; Ganschow and Inscoe; Quinlan and Bernstein), but since those are available elsewhere, it seems useful to focus here, instead, on some larger lessons from the project and on principles that can help shape decisions about how to undertake the peer review of teaching in ways

that will actually improve the things we care about: the quality and conditions of faculty work and the character of student learning.

A first such lesson or principle comes in the form of a gloss on the phrase "The Peer Review of Teaching," which for most faculty means classroom observation in the service of personnel decision-making. What has been clear from the AAHE project, however, is that if what we want is a higher level of teaching and learning, what's needed is a *whole menu* of strategies that faculty can choose among and use to make their work as teachers available to one another—be it to share something they've learned about what helps students learn a hard concept, to be part of an ongoing discussion about a pedagogical issue in the field, to contribute to local decision making about pedagogical and curricular stuff, or, yes, to put their work forward for formal review in the hopes of receiving appropriate recognition and reward.

Thus the peer-review activities of the pilot departments have been deliberately varied—a corrective, as I say, to the view that peer review means exclusively classroom observation, and that its purposes must, by definition, be those of high-stakes personnel decision-making. Rather, faculty teams have undertaken peer-review projects matched to their own local purposes, culture, and needs. In some, for instance, the goal has been to start a conversation about teaching that simply did not exist, and a successful strategy has therefore been the establishment of "teaching circles" and discussion groups. In others, more formal review has been the focus, with faculty collaborating on the design and development of course portfolios that can provide scholarly evidence of teaching for promotion and tenure decisions. A number of departments have also focused on building greater attention to teaching's quality into existing occasions and processes: for instance, asking faculty job candidates, as part of the interview process, to deliver a "pedagogical colloquium" about the teaching of their field. Virtually all have stressed the need to assess student learning, not just teaching.

A second lesson, based on the variety of work undertaken in the pilot departments, is that the relationship between "formative" and "summative" evaluation might be usefully reexamined.

Part of the gospel of evaluation, in teaching and otherwise, is that it's important to distinguish the processes and evidence employed for improvement-oriented, formative purposes from those used for decision-making, summative ones. Nearly all the strategies tried out by pilot departments in the project were originally undertaken with improvement in mind. But one thing a number of faculty have reported is that the same strategies might well be useful for summative/evaluative purposes.

I'm thinking, for instance, of the experience of a faculty member—Peter—in legal studies at the University of Georgia, who, with his departmental colleague—Jere—decided to try out a strategy for interviewing each other's students, part of a larger set of peer- review activities they piloted. The purpose of the interviewing was to try to understand more deeply how students were experiencing their respective courses and teaching, and to gather feedback about possible improvements—which both Peter and Jere were able to make. But in addition, Peter took the initiative of writing a memo to Jere, based on the interviewing experience, summarizing what he thought he had learned about Jere's effectiveness as a teacher. The idea, as he wrote Jere, was that "this might just be of benefit to you" (Hutchings 41)—which in fact it was when Jere was nominated for a teaching award, and he chose to include Peter's memo in his application materials. What was originally private and formative turned out to be useful, as well, in a public, summative context.

This insight is echoed in reports by others. A mathematician from the University of Nebraska developed a course portfolio in order "to know if I'm getting through to the students." He wanted, he said, "more than impressions about this." But he also intends to use the portfolio for an upcoming promotion decision: "I hope to have my portfolio put together and ready to present for review: something that will be comprehensive and data-based in a way that people haven't often seen—something the review committee can sink its teeth into" (Hutchings 57-58). Similarly, a faculty member in English tells of how "teaching development" portfolios constructed by graduate teaching assistants in the composition program at Northern Arizona University later become tools for job- seeking—clearly a summative use.

The point of these examples is not that the formative-summative distinction is one, as they say, without a difference. No doubt about it: it's good to be clear about purposes when undertaking the kinds of processes and practices that can serve to make teaching "community property"; it's good to have ground rules at the outset about what the information will be used for, by whom, and with what, if anything, at stake.

The point, rather, is that when faculty set about making teaching "community property," they develop habits and practices that can, potentially, serve both formative and summative purposes. And though this runs counter to conventional thinking, it shouldn't perhaps be so surprising, since a similar phenomenon is taken fully for granted in research. As research scholars, faculty deliberately seek feedback from the scholarly community; we put our work forward to colleagues for

their insights and contributions and critique. And we do this knowing full well (perhaps even hoping) that some of those same colleagues will judge that work in summative ways when it comes to publication, grants, promotion and tenure. We cross the line between formative improvement and summative evaluation and think little of it. Indeed, doing so is part of what it means to be a member of a scholarly community—be it as researchers or as teachers.

Which brings me to a third lesson, a sort of corollary to the second: that with a little forethought and care, we might, in fact, craft processes for the formal evaluation of teaching that *also* promote improvement.

A case in point is the so-called "pedagogical colloquium" that a number of the pilot departments—the history department at Stanford, for instance—have been experimenting with as part of the process of interviewing faculty job candidates. The colloquium is clearly a summative/evaluative occasion—a very high-stakes one indeed for the job candidates whose teaching abilities are being judged. But it also serves improvement by bringing current faculty into conversation about departmental expectations regarding teaching in ways that are new and improvement-prompting. Indeed, this seemingly secondary consequence of the pedagogical colloquium may be as important as its primary purpose.

A similar dynamic pertains in the use of teaching and course portfolios. Portfolios may be a route to more intellectually credible, authentic evidence for the evaluation of teaching (and this is their original appeal for many faculty), but along the way, the process of their development gets faculty reflecting on their work in powerful new ways—especially when they work in partnership with colleagues who are also developing portfolios.

The punch line here is that though the methods traditionally used to evaluate teaching have not always done much to improve it (and may sometimes even work against improvement), that situation need not be perennial, as Lee Shulman has argued to participants in the AAHE's peer review of teaching project:

> There's a principle that is increasingly employed in discussions of evaluation and assessment today—a principle that we call *"consequential validity."* The point of the principle is that in choosing some form of assessment—of students, of faculty, of whomever—it is not enough todemonstrate that the method is accurate, that it's predictive, that it's fair—though all of those are important criteria. You also must make the argument that the use of a given method of assessment

> or evaluation contributes to the improvement of that which is being evaluated; that the evaluation approach advances the quality of the very enterprise being evaluated. The principle of consequential validity may help us bridge the formative/summative distinction.(3)

In short, Shulman says, "we wish to ensure that whatever we do [to evaluate teaching] contributes to an improvement in the quality of the teaching" (3) and, as many of the faculty participating in the AAHE project would want to add, an improvement in the quality of student learning as well.

Finally, the AAHE project suggests the need to make professional development and improvement be part of what we mean by—and evaluate and reward in—good teaching.

Too often, the kind of teaching that's institutionally valued (though no one says this outright, of course) is teaching without visible defects: students are satisfied, parents do not call the dean's office with complaints, and, in general, instruction is "pulled off" without apparent hitch or glitch. The extreme expression of this ethos is the feeling among faculty on many campuses that seeking assistance with their teaching (say, by visiting the Teaching Center or seeking help from a colleague) is the proverbial kiss of death.

But the peer review of teaching becomes much more powerful if we begin, instead, with a conception of excellent teaching that is not "glitchless," good-enough performance but an ongoing, reflective process aimed always at improvement. Excellent teachers would, by this measure, be those who set out to inquire into their own practice, identifying key issues they want to pursue, posing questions for themselves, exploring alternatives and taking risks, and doing all of this in the company of peers who can offer critique and support. These are the habits of mind we expect, after all, in scholarly work, and we should expect them in teaching as much as in research.

The corollary here is that if excellent teaching entails the deliberate pursuit of improvement, then the deliberate pursuit of improvement (I'll call it "reflective practice") should be an explicit institutional expectation when it comes to summative evaluation. This is, admittedly, a point that can be taken too far: we don't want an evaluation system that rewards a bad teacher for getting a little better more than it rewards achieved excellence. But I at least would argue that we do want to encourage all teachers, not just the novices and the shaky ones, at all stages of their careers, to behave as they do as scholars, seeking new challenges and issues, identifying and solving problems, gathering and using data to guide their practice, consulting with colleagues, and, in

general, contributing to the advancement of good teaching and learning in their own classrooms, and beyond.

What might an evaluation system that values this kind of teaching look like? One answer might be guidelines for portfolio development that call for one entry focused on some problematic dimension of teaching—by which I do not mean a "problem" or personal deficit, but some aspect of teaching the field that is inherently and even universally difficult (e.g., I'm told there's a point about seven weeks into the semester in calculus where large numbers of students fall away) and which therefore needs the attention and thought of teachers willing to go public with their practice.

Even more radically, perhaps, one might imagine criteria for promotion that recognize the possibility and the need for ongoing development by faculty as teachers. At Alverno College, for instance, expectations for teaching differ by rank, with full professors being called upon not only to teach effectively in their own classrooms but to "take leadership" in helping colleagues to teach more effectively and to "influence the professional dialogue" about teaching and learning in higher education—expectations exactly matched to the premise of the project.

The bottom line here is that when it comes to teaching and learning, higher education suffers from a too-low level of ambition. This, I take it, is what Stanford professor William Massy means when he notes the inclination in teaching to "satisfice," to make do, to be content with a certain, not very lofty level of performance and to aim no higher. I would argue that we might counter this low level of ambition by explicitly calling for, evaluating, and rewarding ongoing improvement.

A Word About the Role of Scholarly Societies

From the inception of the AAHE project, the important role of the disciplines has been clear. The activity of the project has been centered, thus far, in a set of pilot departments—originally three on each campus—identified cooperatively by the campuses themselves in order to promote intercampus collaboration by field. The goal has been for historians at Wisconsin to be able to work together with historians at Northwestern, at Georgia, and so forth, a design decision reflecting the fact that for some aspects of teaching, the most relevant peers are scholars from one's own field—first, because teaching history is not teaching chemistry is not teaching engineering, but also, importantly, because the field, not the institution, is for many faculty the primary source of identity and status—and these are exactly what teaching lacks.

With this in mind, we have tried to connect with the relevant scholarly societies to explore ways that the quality and improvement of teaching can be made the subjects of discussion and debate within the community. A number of cooperative ventures have already begun. Articles based on work in the project, authored by faculty participants, have appeared in newsletters and journals from some of the scholarly societies; a number of societies have included sessions on peer review on their annual meeting program. Faculty in one field are thinking about sponsoring a national video-conference on the peer review of teaching. We are eager to help with such efforts (and even have some funding to underwrite them).

The resources listed below will tell you more about how the project has evolved in various disciplinary and campus contexts. For further details, or to have your name added to the project mailing list, please contact: Pam Bender, Program Coordinator, American Association for Higher Education, One Dupont Circle, Suite 360, Washington, DC 20036; e-mail: aaheti.aahe.org; telephone: (202) 293-6440 ext. 56. If you would like to discuss ways in which your scholarly society might be involved, please contact me directly, by e-mail: path@uwyo.edu, or by telephone: (307) 766-4825.

References

Gabaccia, Donna R. "Thinking Globally, Acting Locally: Peer Review of History Teaching at UNC Charlotte." *AHA Perspectives* March 1996: 21-22.

Ganschow, Tom and John Inscoe. "Talking Teaching at the University of Georgia." *AHA Perspectives* April 1996: 29-30.

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

Massy, William F., and Andrea K. Wilger. "Improving Productivity: What Faculty Think About It—And Its Effect on Quality." *Change* 27.4:10-21.

Quinlan, Kathleen and Daniel J. Bernstein, eds. *Innovative Higher Education* 20.4 (1996).

Seldin, Peter. "How Colleges Evaluate Professors: 1983 vs. 1993." *AAHE Bulletin* Oct. 1993: 6-8, 12.

Shulman, Lee S. "The Peer Review of Teaching: A Framework for Action: Three Distinctions." *A Project Workbook.* Washington, DC: American Association for Higher Education, 1995.

University of California. *Report of the University-wide Task Force on Faculty Rewards.* Oakland: University of California, 1991.

Appendix I

AAHE Peer Review of Teaching Project: Participating Campuses

Indiana University-Purdue University Indianapolis
Kent State University
Northwestern University
Stanford University
Syracuse University
Temple University
University of California, Santa Cruz
University of Georgia
University of Michigan
University of Nebraska, Lincoln
University of North Carolina, Charlotte
University of Wisconsin, Madison

Additional campuses will be involved in the next phase of work.

How Evaluations of Teaching Are Used in Personnel Decisions

James England
Temple University

I have been asked to address how two strategies of evaluating teaching, peer review and student evaluation, can come together in the practical setting of personnel decisions.

My experience in higher education has been formed in two quite different circumstances: the small private liberal arts college and the large public research university. In other words, I have had experience at the tails of the distribution of types of institutions in higher education. In most years, one might expect that having information about the tails of a distribution would not be terribly useful in figuring out behavior in the center of the distribution. I will, however, press on, secure in the knowledge that given the character of this presidential election year, your capacity to deduce useful information about the center of a distribution from information about its tails is as acute as it ever will be.

Personnel decisions, more specifically tenure and promotion decisions, are, of course, based on an integrated evaluation of the holy trinity of higher education: teaching, research, and service. While many imagine that the three can be evaluated in isolation from each other, in fact most people recognize that it is some synthetic construction of the three evaluations which eventually produces a complete human being to whom one grants or does not grant tenure. It is also worth noting that the effort put into combining the three in some useful manner increases the probability of making correct personnel decisions. But since the academy has a highly accepted mode of assessing faculty scholarship (blind refereeing or peer review of publications), I will focus my attention on how effective use of the teaching assessment tools of peer review and student evaluation can contribute to a more accurate and more integrated evaluation of a faculty member.

Let me start by expanding on the fairly obvious: that personnel decisions have the institutional mission at their core. Only a small fraction of us have the breadth and quality of performance that make us ideal for any type of institution at any given time. Faculty handbooks notwithstanding, why and how much an institution cares about effective teaching and outstanding scholarship will greatly influence how it goes about judging scholarship and teaching. This is where institutional mission comes into play. Let me cast the issue of institutional mission in terms of the tenure decision, the most significant

personnel decision made in higher education. Institutional mission can often be defined as follows: institutions where effective teaching is a necessary, but not sufficient, condition for granting tenure; and institutions where outstanding scholarship (or the prospect of the same) is a necessary, but not sufficient, condition for granting tenure. The other two possibilities are when either effective teaching or outstanding scholarship is a sufficient condition for being awarded tenure. I will leave out these possibilities from this discussion since they exist only in a caricature of higher education or at institutions where quality is given lip service and where politics rules the day.

At the first type of institution, where effective teaching is a necessary condition for the granting of tenure, typically a high-quality liberal arts college, the quality of (or sometimes the existence of) one's scholarship is important because it is believed to contribute to one remaining an effective teacher over the course of a career of 30 or more years. There are a variety of reasons that guide this thinking:

1. Learning complex material is difficult unless it is taught by someone who is enthusiastically engaged with the material;
2. Effective teaching is mentally exhausting and most of us gain the energy needed from engaging with colleagues in our discipline; and
3. The best way to prevent the "Oracle in the Classroom" syndrome is for faculty members to put their ideas before peers on a regular basis in order to experience the humility brought by a knowledgeable and frank discussion of their scholarship.

While your own list of reasons for including scholarship as a part of teaching evaluation will differ from mine, I think most of us would agree that the effective teacher who is disengaged from the discipline—or who stays current while not actively participating in his or her discipline—is an oxymoron. At a college or university where teaching is central to the institutional mission, scholarship is judged in order to determine whether a faculty member will continue to be an effective teacher over the course of his or her career.

At the second type of institution, where scholarship is a necessary condition for the granting of tenure (typically a research university), scholarship is judged in terms of its importance to the discipline, since the production of knowledge is central to the institution's research mission. While research institutions of high quality subscribe to the liberal arts college's view of the contribution of scholarship to effective teaching, one would imagine that all research institutions would demand both path-breaking research and outstanding teaching. This is

the ideal, but few institutions can afford it. Since effective teaching and the production of significant or field-defining scholarship require a substantial commitment of time, only a few (and, I suspect, a decreasing number of) institutions have the resources to hire a faculty large enough so that individual faculty members can meet both of these standards. It is worth noting that, except at the very few research universities with exceptionally large endowments, the standard for scholarship varies widely by institution and often by department. Most of us, therefore, are in need of some careful institutional soul-searching to align our expectations with our rhetoric, with our financial reality, and with the competing time demands placed on a faculty member's life.

At either type of institution, but keeping in mind that most institutions fall somewhere in between those in my experience, we need to evaluate the quality of teaching and scholarship because they both play a central role in personnel decisions. Depending on the institutional mission, it is a matter of priority or of balance. As I stated above, we do a reasonably good job of evaluating scholarship. We do so because we have a better understanding of why we are interested in scholarship, and we have confidence in the established methods for evaluating scholarship: blind refereeing combined with peer review.

It is the evaluation of teaching that has caused higher education the greatest difficulty and that often results in a skewed profile of a faculty member's teaching effectiveness. I think the reasons for this difficulty can, in part, be explained by the fundamental differences between teaching and research: research is, in the main, a "global" activity, while teaching is a "local" activity.

While the standards for scholarship are derived from the institutional mission, the judgment about whether scholarship has met the standards is discipline-based and, therefore, made outside the university. These external reviews provide us with powerful evidence of the quality of a person's scholarship. At the time of the tenure decision, problems associated with evaluating scholarship usually come from a lack of attention to detail in carrying out the peer review or, more often, from confusion about institutional mission as it relates to scholarship. The former is a problem that we should be able to solve (although I sometimes despair that a large organization may not be able to carry out its procedures with care). The question of institutional mission is one which, as I have said above, must be clear and consistent. That is, we need to be clear in our statements of expectations as they appear in internal memoranda and publications, such as faculty handbooks, and we need to be certain that the public pronouncements of presidents and provosts are consistent with these documents.

I have described teaching as a local activity, meaning that it is evaluated internally. That is, whether one is an effective teacher depends on the depth of knowledge in one's discipline, and also on the expectations of one's departmental colleagues and on one's students' preparation and expectations. Since effective teaching is dependent on the local culture of the institution, any evaluation of teaching must take all three of these factors into account. While recognizing that teaching is, for the most part, a local activity, at too many institutions it has become a private affair.

One result of the privatization of teaching is that it is difficult, if not impossible, to evaluate it reliably. Too often we have cut out colleagues and, in some instances, we have cut out students. Colleagues are often eliminated out of a misguided notion of "academic freedom" or because including colleagues presents us with a difficult and time-consuming process. The elimination of students as a central partner of teaching evaluation causes us to focus on teaching as performance and not on student learning as the reason for teaching. I should also note that involving students in the evaluation of teaching in a thoughtless manner can lead to a perverse notion of effective teaching. We all know of institutions where student evaluations are passed out at the end of every course and in a manner which causes students and faculty alike to conclude that the quality of teaching is directly related to the quality of entertainment. If designed carefully, distributed appropriately, and tabulated thoughtfully, student evaluations of teaching can contribute to the accurate evaluation of an instructor's teaching and can be used to improve the teaching abilities of the instructor being evaluated. Involving peers in the course of reviewing one's teaching can lead to improving its quality and can create a campus climate that supports quality teaching. It can also contribute significant evidence of teaching effectiveness to a personnel file. Involving both, over time, will provide evidence of teaching effectiveness equal in power to the evidence we collect about the quality of scholarship. My fellow contributors to this volume have aptly demonstrated the truth of this statement.

Let me end my remarks by proposing another form of teaching evaluation that is almost unheard of in higher education, but which is becoming widely accepted and often demanded in secondary and elementary education (even if quite controversially so). It is "student outcomes assessment," as it is called in the current public education debate. Only a few colleges and universities have caught on to the value of this form of assessment and have begun to use it in their teaching evaluation. In the case of the departmental major, student outcomes assessment would require departments to:

1. Define standards of performance for persons graduating with their major;
2. Create mechanisms of evaluation to determine whether students have met the standards; and
3. Design curricula and modes of instruction appropriate to these standards.

While student outcomes reflect the quality of teaching not by the quality of a faculty member's input, but by the quality of student learning, the two do not stand in opposition to each other. Student learning results from multiple factors, only one of which is the quality of direct instruction. I am intentionally refraining from supplying details about how to implement student outcomes assessment because each institution will have to develop its own formula for implementation that is appropriate to its own unique circumstances. But I raise this issue here because a critical element of the success of the peer review of teaching project comes from the strong support of professional societies, such as those represented at the ACLS Annual Meeting. At most institutions, but most acutely at research universities, faculty members often take their cues from their professional societies as much as from their home institution. Therefore, support from professional societies can do much to encourage individual institutions to explore student outcomes seriously as an additional measure of faculty and institutional success.

While I am uncertain about how we translate student outcomes assessment into evaluations of individual faculty, I do know that ignoring student learning as demonstrated at an appropriate point in their collegiate careers is contributing to a significant amount of public dismay and even cynicism about our enterprise. For the most part, we have been able to ignore this form of evaluation because the public has accepted the value of higher education with an almost religious fervor. Until quite recently, higher education has not been available to, or required by, a large segment of our society, and we have not had to justify its existence in comparison to other social goods. Times have changed. Now we are being held accountable for what we do by state legislatures and by families who are trying to afford the education we provide. Convincing the public that the education we give is of value and that we are, in fact, educating our students requires us both to describe what we do in clear language and to demonstrate that we are doing it by assessing student outcomes.

Even though student outcomes may not be immediately connected to peer review and student evaluations, I raise the issue here because we need to put it on our collective agendas. It is especially worth

considering as we are now called upon to develop more accurate and integrated ways of evaluating faculty teaching. Using student outcomes goes beyond higher education's conventional understanding of quality teaching and how it is evaluated. We in the academy need to begin looking at the quality of an individual's teaching as it contributes to the effectiveness of the department in which he or she teaches. In turn, a department's effectiveness should then influence individual tenure decisions. If we are to retain the powerful and much-needed public support for, and recognition of, the value of higher education in society, we must be able to demonstrate the value of education through student learning.

ACLS Occasional Papers

1. *A Life of Learning* (1987 Charles Homer Haskins Lecture) by Carl E. Schorske
2. *Perplexing Dreams: Is There a Core Tradition in the Humanities?* by Roger Shattuck
3. *R.M. Lumiansky: Scholar, Teacher, Spokesman for the Humanities*
4. *A Life of Learning* (1988 Charles Homer Haskins Lecture) by John Hope Franklin
5. *Learned Societies and the Evolution of the Disciplines* by Saul B. Cohen, David Bromwich, and George W. Stocking, Jr.
6. *The Humanities in the University: Strategies for the 1990's* by W.R. Connor, Roderick S. French, J. Hillis Miller, Susan Resneck Parr, Merrill D. Peterson, and Margaret B. Wilkerson
7. *Speaking for the Humanities* by George Levine, Peter Brooks, Jonathan Culler, Marjorie Garber, E. Ann Kaplan, and Catharine R. Stimpson
8. *The Agenda for the Humanities and Higher Education for the 21st Century* by Stephen Graubard
9. *A Life of Learning* (1989 Charles Homer Haskins Lecture) by Judith N. Shklar
10. *Viewpoints: Excerpts from the ACLS Conference on The Humanities in the 1990's* by Peter Conn, Thomas Crow, Barbara Jeanne Fields, Ernest S. Frerichs, David Hollinger, Sabine MacCormack, Richard Rorty, and Catharine R. Stimpson
11. *National Task Force on Scholarship and the Public Humanities*
12. *A Life of Learning* (1990 Charles Homer Haskins Lecture) by Paul Oskar Kristeller
13. *The ACLS Comparative Constitutionalism Project: Final Report*
14. *Scholars and Research Libraries in the 21st Century*
15. *Culture's New Frontier: Staking a Common Ground* by Naomi F. Collins
16. *The Improvement of Teaching* by Derek Bok; responses by Sylvia Grider, Francis Oakley, and George Rupp
17. *A Life of Learning* (1991 Charles Homer Haskins Lecture) by Milton Babbitt
18. *Fellowships in the Humanities, 1983-1991* by Douglas Greenberg
19. *A Life of Learning* (1992 Charles Homer Haskins Lecture) by D.W. Meinig
20. *The Humanities in the Schools*

21. *A Life of Learning* (1993 Charles Homer Haskins Lecture) by Annemarie Schimmel
22. *The Limits of Expression in American Intellectual Life* by Kathryn Abrams, W.B. Carnochan, Henry Louis Gates, Jr., and Robert M. O'Neil
23. *Teaching the Humanities: Essays from the ACLS Elementary and Secondary Schools Teacher Curriculum Development Project*
24. *Perspectives on the Humanities and School-Based Curriculum Development* by Sandra Blackman, Stanley Chodorow, Richard Ohmann, Sandra Okura, Sandra Sanchez Purrington, and Robert Stein
25. *A Life of Learning* (1994 Charles Homer Haskins Lecture) by Robert K. Merton
26. *Changes in the Context for Creating Knowledge* by George Keller, Dennis O'Brien, and Susanne Hoeber Rudolph
27. *Rethinking Literary History–Comparatively* by Mario J. Valdés and Linda Hutcheon
28. *The Internationalization of Scholarship and Scholarly Societies*
29. *Poetry In and Out of the Classroom: Essays from the ACLS Elementary and Secondary Schools Teacher Curriculum Development Project*
30. *A Life of Learning* (1995 Charles Homer Haskins Lecture) by Phyllis Pray Bober
31. *Beyond the Academy: A Scholar's Obligations* by George R. Garrison, Arnita A. Jones, Robert Pollack, and Edward W. Said
32. *Scholarship and Teaching: A Matter of Mutual Support* by Francis Oakley
33. *The Professional Evaluation of Teaching* by James England, Pat Hutchings, and Wilbert J. McKeachie