

I This chapter examines the fundamental concepts, principles, and practices that characterize the most effective of contemporary approaches to the evaluation of faculty teaching performance.

Evaluating Teaching Performance

Michael B. Paulsen

There are many comprehensive systems for the evaluation of faculty performance and guidelines for the development of such systems; each includes a substantial component devoted to evaluating faculty teaching performance (Arreola, 2000; Braskamp and Ory, 1994; Cashin, 1996; Centra, 1993; Johnson and Ryan, 2000; Richlin and Manning, 1995; Seldin, 1980, 1999a; Theall and Franklin, 1990). Contributors to this literature agree about key principles that promote effective faculty evaluation (Cashin, 1996). This chapter focuses on three principles: clarifying expectations of and by faculty, identifying the nature and sources of data to be used for evaluation, and clarifying the purposes and uses of evaluation data.

Clarifying Expectations

Clarifying the expectations that institutions and departments have for their faculty and that faculty have for their own performance are central to a successful faculty evaluation system (Arreola, 2000; Braskamp and Ory, 1994; Cashin, 1996; Seldin, 1980, 1999a). Expectations for faculty work responsibilities and outcomes are affected by institutional, departmental, disciplinary, and individual faculty priorities. These expectations also affect the methods, criteria, and the nature and sources of data used to evaluate faculty work. In recent years, both institutional and faculty expectations have begun to change because the nature of faculty work has changed. Redefinitions of faculty roles affect how the teacher role of faculty relates to the other dimensions of faculty work. Understanding long- and short-term changes in the teacher role will help clarify expectations for faculty work as a whole.

Changing Roles and Responsibilities. Teaching competes with other faculty work such as research and service in allocation of faculty time (Austin, 1996; Clark, 1987; Fairweather, 1996). However, several influential reports (for example, Bennett, 1984; Boyer, 1987, 1990; National Institute of Education, 1984) refocused institutional attention and resources on evaluation, improvement, and reward of faculty as teachers. Faculty themselves indicate they value their teaching responsibilities highly. In 1998, 72.8 percent of 33,785 faculty at 378 colleges and universities indicated that their interests were “very heavily in” or “leaning toward” teaching, while only 27.1 percent indicated the same primary interest in research (Sax, Astin, Korn, and Gilmartin, 1999). Faculty interest in teaching persists despite evidence that, across institutional types and different fields of study, faculty who spend the least time on teaching and the most on research receive the highest salaries (Fairweather, 1996).

A new multidimensional view of scholarship that embraces teaching as well as research is changing how people view and value faculty roles and responsibilities (Boyer, 1990; Hutchings and Shulman, 1999; Kreber, 2001; Kreber and Cranton, 2000; Paulsen, 2001; Paulsen and Feldman, 1995b; Rice, 1996). For example, twenty-six professional societies published discipline-specific rationales for restructuring faculty roles and responsibilities to evaluate and reward teaching in ways comparable to research (Diamond and Adam, 2000). In addition, faculty face growing expectations to create student-centered classroom learning environments, focus on active learning, use techniques of classroom assessment and research, and develop pedagogical content knowledge, even though faculty rewards are rarely linked to such teaching innovations (Lazerson, Wagener, and Shumanis, 2000).

Contextual and Flexible Expectations. An institution’s mission and goals provide the framework for most discussions about expectations of and by faculty (Braskamp and Ory, 1994; Cashin, 1996; Johnson and Ryan, 2000), but institutional goals are communicated through departments. Each department has a culture with situation-specific goals, within which faculty expectations are established (Austin, 1996; Cavanagh, 1996). Disciplinary differences affect the relative emphases on teaching and research, goals of undergraduate education, perspectives on curriculum, teaching and students, teaching methods and practices, attitudes toward the improvement of teaching, students’ ratings of teachers, and students’ beliefs about the nature of knowledge and learning (Braxton and Hargens, 1996; Cashin, 1995; Paulsen and Wells, 1998; Smart, Feldman, and Ethington, 2000). Disciplinary differences also affect the nature and construction of pedagogical content knowledge as well as views of what constitutes effective teaching and how it should be evaluated (Hutchings and Shulman, 1999; Shulman, 1993).

Faculty and administrators should discuss expectations, particularly in department units where cultures of institutions and disciplines intersect

(Austin, 1996). An institution's goals are often implicit, and the two groups may have different perceptions of the institution's mission and the relative importance of teaching and research (Gray, Froh, and Diamond, 1992; Gray, Diamond, and Adam, 1996). Braskamp and Ory (1994) recommend, "Given the openness and dynamism of faculty work and careers, it is important that we keep expectations as dynamic and flexible as need be" (p. 59).

Faculty should be actively involved in articulating and negotiating department-specific faculty responsibilities and criteria and methods standards used to evaluate teaching (Arreola, 2000; Braskamp and Ory, 1994; Cashin, 1996; Johnson and Ryan, 2000; Richlin and Manning, 1995; Seldin, 1980, 1999a). Referring to college and university mission statements when formulating department goals will enable participants to understand how department goals relate to institutional goals. Department faculty and the chair should jointly identify the broad categories for faculty roles (teaching, research, service, academic citizenship) and articulate specific responsibilities related to each role (Cashin, 1996).

The question of what constitutes effective teaching must be addressed to identify faculty teaching responsibilities (Cashin, 1996). Careful consideration of the qualities of effective teaching is an especially important undertaking within each department context. No universally accepted definition of effective college teaching exists even though countless attempts have been made to identify the characteristics of effective teaching using a variety of theoretical perspectives and a range of qualitative and quantitative approaches.

Based on extensive consulting experience, Arreola (2000) developed a model for developing a comprehensive evaluation system by systematically collecting faculty input using worksheets and questionnaires. Faculty jointly identify the responsibilities or dimensions of their teaching role (for example, content expertise, instructional design skills, instructional delivery skills), the weight given to each dimension, the sources of information about performance on each dimension (for example, students, peers, self, chair), and the impact weight of each source for each dimension. They also identify the tool for collecting data from each source for each dimension (for example, questionnaire, review of course materials, interview, and so forth), a common rating scale, the method of computing composite weighted ratings for teaching dimensions (such as 1.25 for content expertise, 1.0 for instructional design, and 1.25 for instructional delivery), the computation of a composite weighted rating for the teaching role ($1.25 + 1.0 + 1.25 = 3.5$), and a minimum overall composite rating (such as 2.5).

Every faculty member should meet regularly (annually at least) with the department chair to discuss and agree on the nature of that individual's responsibilities and the methods, sources, and criteria that will be used to evaluate teaching performance (Cashin, 1996; Seldin, 1980). Faculty members should be able to focus their efforts on those activities that best match their

own interests, skills, and experience by negotiating with their colleagues how they can best use their talents to contribute to the collective work of their department (Wergin and Swingen, 2000).

The Nature and Sources of Data for the Evaluation of Teaching

Multiple sources and types of data should be used to evaluate teaching. The most common sources of data are students, peers, and teachers themselves (Centra, 1993; Paulsen and Feldman, 1995a; Seldin, 1999b; Theall and Franklin, 1990).

Student Ratings. Quantitative student ratings of teaching are used more than any other method to evaluate teaching performance (Cashin, 1999; Seldin, 1999b). Student ratings play a dominant role in the operational definition of what constitutes effective teaching. Components of effective teaching identified from analysis of student ratings include six common dimensions of skill, rapport, structure, difficulty, interaction, and feedback (Cohen, 1981). Other scholars have identified from nine (Marsh, 1984) to as many as twenty-eight dimensions (Feldman, 1997).

Even though student ratings are widely used and despite the large volume of research demonstrating their validity and reliability, faculty express concerns about their meaningfulness and appropriateness. Franklin and Theall (1989) found from their survey of more than six hundred faculty and administrators at three colleges that those with greater knowledge and awareness about research on student ratings had more favorable attitudes toward the use of student ratings in teaching evaluation than those with less knowledge.

The reliability of student ratings is generally robust (Cashin, 1995; Feldman, 1977; Marsh and Dunkin, 1997). Reliability coefficients for consistency (interrater agreement) vary according to the number of students surveyed but are about .70 or higher when more than ten raters are surveyed on well-known rating forms such as the Student Instructional Report (SIR) (Centra, 1993), the Student Evaluation of Educational Quality (SEEQ) (Marsh, 1984), and the Instructional Development and Effectiveness Assessment (IDEA) (Cashin, 1995). Reliability coefficients for stability (agreement of ratings over time) are also impressive, with average correlations of .83 between student ratings at semester's end their ratings one or more years later (Marsh and Dunkin, 1997). Reliability estimates that assess the extent to which student ratings of an instructor generalize across different courses or different offerings of the same course produce coefficients of .61 and .72, respectively (Marsh, 1984). In combination, these findings indicate that for summative purposes, ratings for an instructor should be collected from an adequate number of students and should cover different courses and years (Cashin, 1999; Centra, 1993; Marsh, 1984).

The validity of student ratings is assessed by the extent to which they measure a generally agreed-upon indicator of teaching effectiveness; correlate with ratings assigned by the teachers themselves, their colleagues, administrators or alumni; or agree with qualitative student evaluations (Braskamp and Ory, 1994; Cashin, 1995; Centra, 1993; Feldman, 1989a, 1989b, 1997; Marsh and Dunkin, 1997).

Metanalyses of student ratings in for a large number of multisection courses resulted in moderate (over .30) to strong (over .50) correlations between ratings on separate dimensions and global items and student performance on common final examinations (Cohen, 1981; Feldman, 1989a). Another metaanalysis (Feldman, 1989b) resulted in average correlations between student ratings and the ratings of the following other groups: alumni (.69), instructors themselves (.29), colleagues (.55), administrators (.39), and external, trained observers who had either visited instructors' classrooms or viewed videotapes of their teaching (.50). Qualitative (written or group interview) evaluations by students are highly correlated with their quantitative ratings (Braskamp and Ory, 1994). In combination, these findings provide general support for the validity of student ratings in the evaluation of teaching.

Cashin's comprehensive matrix (1989, tab. 1) indicates which sources provide data appropriate for evaluating various aspects of faculty teaching performance, thereby addressing issues of face validity. He identifies seven areas that, in combination, capture the complex concept of teaching. Four of these are appropriate for students to evaluate: delivery of instruction, assessment of instruction, availability to students, and administrative requirements.

Possible biases in student ratings must be considered, especially when ratings are to be used for summative purposes. "Bias" may be present when instructor, student, course, or administrative variables are correlated with student ratings but are "not related to teaching effectiveness" (Cashin, 1995, p. 4). Research has identified few variables that meet these conditions. Two distinctions are important. The first is the distinction between variables that are or are not related to student ratings in a way that could lead to possible bias. The second distinction is among variables related to student ratings and either appropriately related to teaching effectiveness or not.

Variables related to student ratings but not teaching effectiveness may require statistical control. They include student, course, and administrative variables. One student variable requiring control is motivation for taking courses. Students taking a course as an elective tend to give higher ratings than those taking a course as a requirement (Feldman, 1978). A student's expected grade is also correlated with student ratings of instructors (Feldman, 1976). Expert evaluators disagree about whether grading leniency in conjunction with workload does or does not bias student ratings (Greenwald and Gillmore, 1997a, 1997b; Marsh, 1984; Marsh and Dunkin, 1997; Marsh and Roche, 2000).

Several course variables are related to student ratings. Students tend to rate graduate and upper-division courses higher than undergraduate and lower-division courses, respectively. Students rate courses in the arts and humanities somewhat higher than in the social sciences, which in turn, are rated higher than math and science courses (Cashin, 1995; Feldman, 1978; Marsh, 1984; Marsh and Dunkin, 1997). If faculty and administrators observe that students rate differently by level of course, academic field, or motivation for taking the course and if they suspect that such differences may be due to differences in the characteristics of the students or courses and not to differences in the effectiveness of the teachers, then normative or comparative groups should be established to promote greater fairness in the comparison of ratings (Cashin, 1995, 1999). Finally, student ratings tend to be higher in relation to the following administrative factors: when the instructor is present, when students know the purpose is for personnel decisions, and when forms are not anonymous (Braskamp and Ory, 1994; Centra, 1993). These issues can be controlled by using standardized instructions to students regarding the purposes of the ratings, asking students not to sign their names, and requiring instructors to leave the classroom while forms are completed (Cashin 1995, 1999).

Peer Review of Teaching. Although many experts agree that students are qualified to assess many aspects of classroom teaching (for example, clarity of presentation, interpersonal rapport with students, concern for students progress), they also assert that for some aspects of teaching (mastery of content, course goals, course organization and materials), only peers have the substantive expertise required for meaningful evaluation (Cashin, 1989; Chism, 1999; Hutchings, 1996b). In short, peer review brings content-based contextuality to evaluation of teaching.

Specialists in teaching and its evaluation also agree that the work of an individual faculty member is valued more when it has been subjected to rigorous peer review (Cavanagh, 1996; Chism, 1999; Diamond and Adam, 2000; Hutchings, 1996a, 1996b). Therefore, research is more highly valued than teaching (Boyer, 1990). Faculty expect public review of the methods and products of their research. In contrast, methods and products of teaching are rarely discussed or shared with peers. Just as the quality of research improves due to dialogue and debate among disciplinary peers, so would the quality of teaching benefit from similar opportunities (Boyer, 1990; Chism, 1999; Hutchings, 1996a, 1996b).

Proponents of peer review of teaching acknowledge a set of key issues and concerns, including privacy of the reviewed, needs of the reviewer, and reliability and validity of the ratings. What goes on in the classroom has traditionally been between teachers and their students, not between teachers and their peers. Peer review challenges norms of privacy by opening doors to classrooms and making teaching a public act (Chism, 1999; Hutchings, 1996a). Ending “pedagogical solitude” may be uncomfortable for many

faculty (Shulman, 1993). Yet faculty are sharing many stories of successful experiments with peer collaboration and peer review (Hutchings, 1996a; Langsam and Dubois, 1996; Nordstrom, 1995; Quinlan, 1996). These changes bring increasing opportunities for new faculty to be mentored in ways that socialize them to peer collaboration and review (Hutchings and Shulman, 1999), and they may promote a culture of collaboration and community surrounding teaching (Cavanagh, 1996; Hutchings, 1996a, 1996b).

Reviewers' concerns must also be considered (Cavanagh, 1996; Chism, 1999; DeZure, 1999; Hutchings, 1996a, 1996b; Seldin, 1980). Without careful planning, peer reviewers could be placed in awkward situations when asked to judge a colleague, wrestle with issues of confidentiality, risk lack of anonymity, assess the strengths or weaknesses of a senior colleague, worry about potential ambiguous legal issues, and devote time and energy to matters that they may not perceive to be part of their job (Centra, 1993; Chism, 1999; Hutchings, 1996a, 1996b).

The reliability and validity of peer ratings of teaching are not as well established as they are for student ratings. Classroom observations of teaching have been used in a growing number of institutions (Seldin, 1999b). Research has indicated that, in the absence of either sound training or adequate numbers of observers, peer ratings based *solely* on classroom observation are not generally reliable (Centra, 1993). Questions of validity arise about whether the presence of an obtrusive observer might alter classroom behavior (Cohen and McKeachie, 1980). But there is general consensus that training in the observation of classroom teaching and that increasing the number of observers and the number of visits they make to each class would, in combination, increase the reliability of peer classroom observation to acceptable levels (Braskamp and Ory, 1994; Centra, 1993; Chism, 1999; DeZure, 1999). Departments using peer observation to evaluate classroom teaching should follow sound procedures in selecting and training observers; identifying the number of observers and number and length of classroom visits; collecting data to use in assessing the reliability and validity of observers and observations; establishing guidelines, criteria, and standards for observation; developing forms and methods for making observations; and preparing the report of the observations (Chism, 1999; DeZure, 1999).

Studies of general peer ratings of overall teaching effectiveness have produced reliability estimates ranging from .64 to .86 (Cohen and McKeachie, 1980). Kremer (1990) found that reliability across all peer raters was only .50, but when only those raters who indicated that they had high confidence in their rating were considered, reliability estimates increased to .82. Correlations with student ratings have ranged from .62 to .87. Feldman's metaanalysis (1989b), based on some studies that did and some that did not include classroom observations, resulted in an average correlation with student ratings of .55. The correlations between student and peer ratings may be somewhat overstated because one of the likely bases for peers' ratings

were previous ratings of students available to them—that is, the two sets of ratings may not be entirely independent (Cohen and McKeachie, 1980; Feldman, 1989b; Marsh and Dunkin, 1997).

What are peer reviewers best qualified to evaluate? Peer review should be used to provide data on aspects of teaching effectiveness for which faculty peers are the best available source of information (Arreola and Aleamoni, 1990; French-Lazovik, 1981), including expertise in the subject matter and discipline-specific aspects of instructional design and pedagogy (Arreola, 2000; Chism, 1999; Shulman, 1993). Five areas appropriate for peer review are subject matter mastery, curriculum development, course design, delivery of instruction, and assessment of instruction (Cashin, 1989). Only peers can evaluate the first three, whereas both peers and students can evaluate the last two.

Self-Evaluation or Report: Peer Review of the Teaching Portfolio.

Although self-evaluations by teachers lack the validity and objectivity necessary for summative evaluation (Centra, 1993), support is growing for the use of teaching portfolios with data supplied by the instructor (Arreola, 2000; Braskamp and Ory, 1994; Centra, 2000; Chism, 1999; Seldin, 1993). “Course syllabi and exams” and “self-evaluation or report” were among the fastest-growing sources of data used in evaluating teaching performance between 1988 and 1998 (Seldin, 1999b, p. 14). The expanding use of these data sources is consistent with the nationwide increase in the use of peer review of portfolios to evaluate faculty teaching performance.

Research on reliability of peer review of teaching portfolios appears promising. In one study, faculty elected a six-member executive committee to evaluate all faculty dossiers in the areas of research, teaching, and service (Root, 1987). The composite reliability coefficients for the six raters were .97, .90, and .90 for research, teaching, and service, respectively. In a study of peer review of teaching portfolios for summative purposes at a community college, faculty wrote personal statements and provided descriptions, examples, and other documentation of their teaching effectiveness on thirteen aspects of teaching arranged into three categories: motivational skills, interpersonal skills, and intellectual skills (Centra, 1993, 1994). Each portfolio was evaluated by a dean, one peer selected by the instructor, and another peer designated by the dean. The ratings by the dean and the peer designated by the dean (peer B) were significantly correlated with each other and with student ratings; however, the ratings of the peer selected by the instructor (peer A) were not significantly related to those of the other raters.

Several steps can enhance the reliability and validity of peer ratings of teaching portfolios in summative evaluation (Centra, 1993, 1994, 2000; Root, 1987). First, portfolios should include a broad range of work samples and related information to document various aspects of teaching performance. Second, peer reviewers should receive training that includes opportunities to discuss methods, criteria, and standards for assessment using portfolios

that have previously been rated high or low. Third, peers' objectivity as reviewers will be enhanced if they are not being currently evaluated and if they are selected by the unit head, randomly selected, or elected to a peer committee on teaching. Fourth, a minimum of three and a maximum of six peer reviewers should be used.

Reliability of peer ratings of portfolios would also be enhanced if a set of mandated items were included in every portfolio. Seldin (1993) recommends the following items: a reflective statement about the instructor's teaching approach, three years' of summaries of student ratings, three years' of syllabi for all courses taught, innovative course materials, and evidence of activities to improve one's teaching. Chism's sourcebook on peer review provides other models and detailed guidance (Chism, 1999).

Purposes and Uses of Evaluation Data

Evidence on teaching effectiveness can be collected for two uses—formative and summative (Braskamp and Ory, 1994; Centra, 1993; Marsh and Dunkin, 1997; Paulsen and Feldman, 1995a; Theall and Franklin, 1990). The purpose of formative evaluation is to provide informative feedback to assist faculty in improving the effectiveness of their teaching. The purpose of summative evaluation is to provide information to assist department chairs, faculty committees, and deans in making personnel decisions related to hiring, renewing or terminating faculty, awarding tenure, promotion, and merit pay increases.

To address these different purposes effectively, different types of information may be needed from the evaluation system (Abrami and d'Apollonia, 1990; Arreola and Aleamoni, 1990; Cashin, 1999; Theall and Franklin, 1990). For developmental purposes, the evaluation system should generate regularly collected and detailed diagnostic data that can be confidentially provided to individual teachers to help them identify strengths and weaknesses in their teaching behavior, establish priorities, and plan strategies for teaching improvement. Detailed diagnostic data may not be essential or appropriate for summative purposes. Instead, evaluation may be based on summary data from multiple sources of a teacher's overall teaching performance in more than one course over an extended period of time. However, the relationship between formative and summative evaluation is no less important than the distinction between them. If evaluation data and procedures are used for formative (developmental) purposes prior to being used for summative purposes (judgment), faculty have opportunities to become more familiar with the nature of the data, methods, and criteria that will be for subsequent summative evaluation of their teaching. As a result, they can strive to improve their performance before it is formally assessed (Centra, 1993). Hutchings (1996a) makes a strong argument for bridging the summative-formative distinction. When faculty are successful in making teaching community property and when they construct a culture of collaboration and peer review

around teaching, they will have attitudes and perform actions that could serve both summative and formative purposes well. This bridging already happens with research (Hutchings, 1996a). Even advanced and experienced researchers deliberately seek formative evaluation (informative feedback) from peers to help improve the quality of their research work. They seek feedback, knowing very well that their peers are also likely to judge their research performance in summative ways, such as reviewing abstracts for conference presentations, manuscripts submitted for publication, or proposals for funded research or such as serving as external reviewers for tenure, promotion, or awards. In their roles as researchers, faculty bridge the formative-summative distinction without hesitation or concern. Perhaps soon, the distinction between summative and formative evaluation of teaching will no longer be useful or meaningful, because teaching will have become community property, just as research has been for so many years.

Summary

This chapter examined concepts, principles, and practices of effective contemporary approaches for evaluating faculty teaching performance. The chapter included elements of comprehensive systems for the evaluation of teaching performance, including faculty roles and responsibilities, criteria and methods for evaluating faculty performance. Next, the sources, types, reliability and validity of data used for evaluation, including student ratings, peer review, self-report and portfolios, were examined in some depth. The roles of rewards, disciplinary perspectives, and institutional teaching cultures in the development of effective teaching evaluation systems were considered from a variety of perspectives. Finally, the distinctions and relationships between formative and summative evaluation were discussed from philosophical, conceptual, and practical perspectives.

References

- Abrami, P. C., and d'Apollonia, S. "The Dimensionality of Ratings and Their Use in Personnel Decisions." In M. Theall and J. Franklin (eds.), *Student Ratings of Instruction: Issues for Improving Practice*. New Directions for Teaching and Learning, no. 43. San Francisco: Jossey-Bass, 1990.
- Aleamoni, L. M. "Student Rating Myths Versus Research Facts: An Update." *Journal of Personnel Evaluation in Education*, 1999, 13(2), 153–166.
- Arreola, R. *Developing a Comprehensive Faculty Evaluation System: A Handbook for College Faculty and Administrators on Designing and Operating a Comprehensive Faculty Evaluation System*. Bolton, Mass.: Anker, 2000.
- Arreola, R., and Aleamoni, L. "Practical Issues in Designing and Operating a Faculty Evaluation System." In M. Theall and J. Franklin (eds.), *Student Ratings of Instruction: Issues for Improving Practice*. New Directions for Teaching and Learning, no. 43. San Francisco: Jossey-Bass, 1990.
- Austin, A. "Institutional and Departmental Cultures: The Relationship Between Teaching and Research." In J. Braxton (ed.), *Faculty Teaching and Research: Is There*

- a Conflict?* New Directions for Institutional Research, no. 90. San Francisco: Jossey-Bass, 1996.
- Batista, E. "The Place of Colleague Evaluation in the Appraisal of College Teaching: A Review of the Literature." *Research in Higher Education*, 1976, 4, 257–271.
- Bennett, W. J. *To Reclaim a Legacy*. Washington, D.C.: National Endowment for the Humanities, 1984.
- Boyer, E. L. *College: The Undergraduate Experience in America*. New York: Harper and Row, 1987.
- Boyer, E. L. *Scholarship Reconsidered: Priorities of the Professoriate*. Princeton, N.J.: The Carnegie Foundation for the Advancement of Teaching, 1990.
- Braskamp, L. A., and Ory, J. C. *Assessing Faculty Work: Enhancing Individual and Institutional Performance*. San Francisco: Jossey-Bass, 1994.
- Braxton, J., and Hargens, L. "Variation Among Academic Disciplines: Analytical Frameworks and Research." In J. Smart (ed.), *Higher Education: Handbook of Theory and Research*. Vol. 11. New York: Agathon Press, 1996.
- Cashin, W. E. *Defining and Evaluating College Teaching*. Idea Paper, no. 21. Manhattan, Kans.: Center for Faculty Evaluation and Faculty Development, Kansas State University, 1989.
- Cashin, W. E. *Student Ratings of Teaching: The Research Revisited*. Idea Paper, no. 32. Manhattan, Kans.: Center for Faculty Evaluation and Faculty Development, Kansas State University, 1995.
- Cashin, W. E. *Developing an Effective Faculty Evaluation System*. Idea Paper, no. 33. Manhattan, Kans.: Center for Faculty Evaluation and Faculty Development, Kansas State University, 1996.
- Cashin, W. E. "Student Ratings of Teaching: Uses and Misuses." In P. Seldin (ed.), *Current Practices in Evaluating Teaching: A Practical Guide to Improved Faculty Performance and Promotion/Tenure Decisions*. Bolton, Mass.: Anker, 1999.
- Cavanagh, R. R. "Summative and Formative Evaluation in the Faculty Peer Review of Teaching." *Innovative Higher Education*, 1996, 20(4), 235–240.
- Centra, J. A. *Reflective Faculty Evaluation: Enhancing Teaching and Determining Faculty Effectiveness*. San Francisco: Jossey-Bass, 1993.
- Centra, J. A. "The Use of the Teaching Portfolio and Student Evaluations for Summative Evaluation." *The Journal of Higher Education*, 1994, 65(5), 555–570.
- Centra, J. A. "Evaluating the Teaching Portfolio: A Role for Colleagues." In K. E. Ryan (ed.), *Evaluating Teaching in Higher Education: A Vision for the Future*. New Directions for Teaching and Learning, no. 83. San Francisco: Jossey-Bass, 2000.
- Chism, N. *Peer Review of Teaching: A Sourcebook*. Bolton, Mass.: Anker, 1999.
- Clark, B. R. *The Academic Life: Small Worlds, Different Worlds*. Princeton, N.J.: The Carnegie Foundation for the Advancement of Teaching, 1987.
- Cohen, P. A. "Student Ratings of Instruction and Student Achievement: A Meta-analysis of Multisection Validity Studies." *Review of Educational Research*, 1981, 51(3), 281–309.
- Cohen, P. A., and McKeachie, W. J. "The Role of Colleagues in the Evaluation of College Teaching." *Improving College and University Teaching*, 1980, 28(4), 147–154.
- DeZure, D. "Evaluating Teaching Through Peer Classroom Observation." In P. Seldin (ed.), *Current Practices in Evaluating Teaching: A Practical Guide to Improved Faculty Performance and Promotion/Tenure Decisions*. Bolton, Mass.: Anker, 1999.
- Diamond, R. M., and Adam, B. E. (eds.). *The Disciplines Speak II: More Statements on Rewarding the Scholarly, Professional, and Creative Work of Faculty*. Washington, D.C.: American Association for Higher Education, 2000.
- Fairweather, J. S. *Faculty Work and Public Trust: Restoring the Value of Teaching and Public Service in American Academic Life*. Boston: Allyn & Bacon, 1996.
- Feldman, K. A. "Grades and College Students' Evaluations of Their Courses and Teachers." *Research in Higher Education*, 1976, 4, 69–111.

- Feldman, K. A. "Consistency and Variability Among College Students in Rating Their Teachers and Courses: A Review and Analysis." *Research in Higher Education*, 1977, 6(3), 223–274.
- Feldman, K. A. "Course Characteristics and College Students' Ratings of Their Teachers." *Research in Higher Education*, 1978, 9, 199–242.
- Feldman, K. A. "The Association Between Student Ratings of Specific Instructional Dimensions and Student Achievement." *Research in Higher Education*, 1989a, 30(6), 583–645.
- Feldman, K. A. "Instructional Effectiveness of College Teachers as Judged by Teachers Themselves, Current and Former Students, Colleagues, Administrators, and External (Neutral) Observers." *Research in Higher Education*, 1989b, 30(2), 113–135.
- Feldman, K. A. "Identifying Exemplary Teachers and Teaching: Evidence from Student Ratings." In R. Perry and J. Smart (eds.), *Effective Teaching in Higher Education: Research and Practice*. New York: Agathon Press, 1997.
- Franklin, J., and Theall, M. "Who Reads Ratings: Knowledge, Attitudes, and Practices of Users of Student Ratings of Instruction." Paper presented at the annual meeting of the American Educational Research Association, San Francisco, Apr. 1989.
- French-Lazovik, G. "Peer Review: Documentary Evidence in the Evaluation of Teaching." In J. Millman (ed.), *Handbook of Teacher Evaluation*. Beverly Hills: Sage, 1981.
- Gray, P. J., Diamond, R. M., and Adam, B. E. *A National Study on the Relative Importance of Research and Undergraduate Teaching at Colleges and Universities*. Syracuse, N.Y.: Center for Instructional Development, Syracuse University, 1996.
- Gray, P. J., Froh, R. C., and Diamond, R. M. *A National Study of Research Universities on the Balance Between Research and Undergraduate Teaching*. Syracuse, N.Y.: Center for Instructional Development, Syracuse University, 1992.
- Greenwald, A. G., and Gillmore, G. M. "Grading Leniency Is a Removable Contaminant of Student Ratings." *American Psychologist*, 1997a, 52(11), 1209–1217.
- Greenwald, A. G., and Gillmore, G. M. "No Pain, No Gain? The Importance of Measuring Course Workload in Student Ratings of Instruction." *Journal of Educational Psychology*, 1997b, 89(4), 743–751.
- Hutchings, P. *Making Teaching Community Property: A Menu for Peer Collaboration and Peer Review*. Washington, D.C.: American Association for Higher Education, 1996a.
- Hutchings, P. "The Peer Review of Teaching: Progress, Issues and Prospects." *Innovative Higher Education*, 1996b, 20(4), 221–234.
- Hutchings, P., and Shulman, L. "The Scholarship of Teaching: New Elaborations, New Developments." *Change*, 1999, 31(5), 11–15.
- Johnson, T., and Ryan, K. "A Comprehensive Approach to the Evaluation of College Teaching." In K. E. Ryan (ed.), *Evaluating Teaching in Higher Education: A Vision for the Future*. New Directions for Teaching and Learning, no. 83. San Francisco: Jossey-Bass, 2000.
- Kreber, C. (ed.). *Scholarship Revisited: Defining and Implementing the Scholarship of Teaching*. New Directions for Teaching and Learning, no. 86. San Francisco: Jossey-Bass, 2001.
- Kreber, C., and Cranton, P. "Exploring the Scholarship of Teaching." *Journal of Higher Education*, 2000, 71(4), 476–495.
- Kremer, J. F. "Construct Validity of Multiple Measures in Teaching, Research, and Service and Reliability of Peer Ratings." *Journal of Educational Psychology*, 1990, 82(2), 213–218.
- Langsam, D. M., and Dubois, P. L. "Can Nightmares Become Sweet Dreams? Peer Review in the Wake of a Systemwide Administrative Mandate." *Innovative Higher Education*, 1996, 20(4), 249–259.
- Lazerson, M., Wagener, U., and Shumanis, N. "Teaching and Learning in Higher Education, 1980–2000." *Change*, 2000, 32(3), 13–19.

- Marsh, H. W. "Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases, and Utility." *Journal of Educational Psychology*, 1984, 76, 707–754.
- Marsh, H. W., and Dunkin, M. J. "Students' Evaluations of University Teaching: A Multidimensional Perspective." In R. Perry and J. Smart (eds.), *Effective Teaching in Higher Education: Research and Practice*. New York: Agathon Press, 1997.
- Marsh, H. W., and Roche, L. A. "Effects of Grading Leniency and Low Workload on Students' Evaluations of Teaching: Popular Myth, Bias, Validity, or Innocent Bystanders?" *Journal of Educational Psychology*, 2000, 92(1), 202–228.
- National Institute of Education. *Involvement in Learning: Realizing the Potential of American Higher Education*. Washington, D.C.: U.S. Government Printing Office, 1984.
- Nordstrom, K. "Multiple-Purpose Use of a Peer Review of Course Instruction Program in a Multidisciplinary University Department." *Journal on Excellence in College Teaching*, 1995, 6(3), 125–144.
- Paulsen, M. B. "The Relation Between Research and the Scholarship of Teaching." In C. Kreber, (ed.), *Scholarship Revisited: Defining and Implementing the Scholarship of Teaching*. New Directions for Teaching and Learning, no. 86. San Francisco: Jossey-Bass, 2001.
- Paulsen, M. B., and Feldman, K. A. *Taking Teaching Seriously: Meeting the Challenge of Instructional Improvement*. ASHE-ERIC Higher Education Report, no. 6. Washington, D.C.: Association for the Study of Higher Education, 1995a.
- Paulsen, M. B., and Feldman, K. A. "Toward a Reconceptualization of Scholarship: A Human Action System with Functional Imperatives." *The Journal of Higher Education*, 1995b, 66(6), 615–640.
- Paulsen, M. B., and Wells, C. "Domain Differences in the Epistemological Beliefs of College Students." *Research in Higher Education*, 1998, 39(4), 365–384.
- Quinlan, K. M. "Involving Peers in the Evaluation and Improvement of Teaching: A Menu of Strategies." *Innovative Higher Education*, 1996, 20(4), 299–307.
- Rice, R. E. *Making a Place for the New American Scholar*. Washington, D.C.: American Association for Higher Education, 1996.
- Richlin, L., and Manning, B. "Evaluating College and University Teaching: Principles and Decisions for Designing a Workable System." *Journal on Excellence in College Teaching*, 1995, 6(3), 3–15.
- Root, L. S. "Faculty Evaluation: Reliability of Peer Assessments of Research, Teaching, and Service." *Research in Higher Education*, 1987, 26(1), 71–84.
- Sax, L., Astin, A., Korn, W., and Gilmartin, S. *The American College Teacher*. Los Angeles: Higher Education Research Institute, UCLA, 1999.
- Seldin, P. *Successful Faculty Evaluation Programs*. Crugers, N.Y.: Coventry Press, 1980.
- Seldin, P. *Successful Use of Teaching Portfolios*. Bolton, Mass.: Anker, 1993.
- Seldin, P. "Building Successful Teaching Evaluation Programs." In P. Seldin (ed.), *Current Practices in Evaluating Teaching: A Practical Guide to Improved Faculty Performance and Promotion/Tenure Decisions*. Bolton, Mass.: Anker, 1999a.
- Seldin, P. "Current Practices—Good and Bad—Nationally." In P. Seldin (ed.), *Current Practices in Evaluating Teaching: A Practical Guide to Improved Faculty Performance and Promotion/Tenure Decisions*. Bolton, Mass.: Anker, 1999b.
- Shulman, L. S. "Teaching as Community Property: Putting an End to Pedagogical Solitude." *Change*, 1993, 25(6), 6–7.
- Smart, J., Feldman, K., and Ethington, C. *Academic Disciplines: Holland's Theory and the Study of College Students and Faculty*. Nashville, Tenn.: Vanderbilt University Press, 2000.
- Theall, M., and Franklin, J. "Student Ratings in the Context of Complex Evaluation Systems." In M. Theall and J. Franklin (eds.), *Student Ratings of Instruction: Issues for Improving Practice*. New Directions for Teaching and Learning, no. 43. San Francisco: Jossey-Bass, 1990.

- Theall, M., and Franklin, J. (eds.). *Effective Practices for Improving Teaching*. New Directions for Teaching and Learning, no. 48. San Francisco: Jossey-Bass, 1991.
- Weimer, M. *Improving College Teaching*. San Francisco: Jossey-Bass, 1990.
- Wergin, J. F., and Swingen, J. N. *Departmental Assessment: How Some Campuses Are Effectively Evaluating the Collective Work of Faculty*. Washington, D.C.: American Association for Higher Education, 2000.

MICHAEL B. PAULSEN is professor of education and coordinator of graduate studies in higher education in the department of educational leadership, counseling, and foundations at the University of New Orleans.