International Journal of Teaching and Learning in Higher Education 2005, Volume 17, Number 1, 48-62
http://www.isetl.org/ijtlhe/ ISSN 1812-9129

Survey of 12 Strategies to Measure Teaching Effectiveness

Ronald A. Berk
Johns Hopkins University, USA

Twelve potential sources of evidence to measure teaching effectiveness are critically reviewed: (a) student ratings, (b) peer ratings, (c) self-evaluation, (d) videos, (e) student interviews, (f) alumni ratings, (g) employer ratings, (h) administrator ratings, (i) teaching scholarship, (j) teaching awards, (k) learning outcome measures, and (l) teaching portfolios. National standards are presented to guide the definition and measurement of effective teaching. A unified conceptualization of teaching effectiveness is proposed to use multiple sources of evidence, such as student ratings, peer ratings, and self-evaluation, to provide an accurate and reliable base for formative and summative decisions. Multiple sources build on the strengths of all sources, while compensating for the weaknesses in any single source. This triangulation of sources is recommended in view of the complexity of measuring the act of teaching and the variety of direct and indirect sources and tools used to produce the evidence.

Yup, that's what I typed: *12*. A virtual smorgasbord of data sources awaits you. How many can you name other than student ratings? How many are currently being used in your department? That's what I thought. This is your lucky page. By the time you finish this article, your toughest decision will be (Are you ready? Isn't this exciting?): Should I slog through the other *IJTLHE* articles? WROOONG! It's: Which sources should I use?

Teaching Effectiveness: Defining the Construct

Why is measuring *teaching effectiveness* so important? Because the evidence produced is used for major decisions about our future in academe. There are two types of decisions: *formative,* which uses the evidence to improve and shape the quality of our teaching, and *summative*, which uses the evidence to "sum up" our overall performance or status to decide about our annual merit pay, promotion, and tenure. The former involves decisions to improve teaching; the latter consists of personnel decisions. As faculty, we make formative decisions to plan and revise our teaching semester after semester. Summative decisions are final and they are rendered by administrators or colleagues at different points in time to determine whether we have a future. These decisions have an impact on the quality of our professional life. The various sources of evidence for teaching effectiveness may be employed for either formative or summative decisions or both.

National Standards

There are national standards for how teaching effectiveness or performance should be measured—the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME Joint Committee on Standards, 1999). They can guide the development of the measurement tools, the technical analysis of the results, and the reporting and interpretation of the evidence for decision making.

The *Standards* address WHAT is measured and then HOW to measure it: WHAT – The content of any tool, such as a student or peer rating scale, requires a thorough and explicit definition of the knowledge, skills, and abilities (KSAs), and other characteristics and behaviors that describe the job of "effective teaching" (see Standards 14.8–14.10). HOW – The data from a rating scale or other tool that is based on the systematic collection of opinions or decisions by raters, observers, or judges hinge on their expertise, qualifications, and experience (see Standard 1.7).

Student and peer direct observations of WHAT they see in the classroom furnish the foundation for their ratings. However, other sources, such as student outcome data and publications on innovative teaching strategies, are indirect, from which teaching effectiveness is inferred. These different data sources vary considerably in how they measure the WHAT. We need to be able to carefully discriminate among all available sources.

Beyond Student Ratings

Historically, student ratings have dominated as the primary measure of teaching effectiveness for the past 30 years (Seldin, 1999a). However, over the past decade there has been a trend toward augmenting those ratings with other data sources of teaching performance. Such sources can serve to broaden and deepen the evidence base used to evaluate courses and assess the quality of teaching (Arreola, 2000; Braskamp & Ory, 1994; Knapper & Cranton, 2001; Seldin & Associates, 1999).

Several *comprehensive* models of faculty evaluation have been proposed (Arreola, 2000; Braskamp & Ory, 1994; Centra, 1999; Keig &

Waggoner, 1994; Romberg, 1985; Soderberg, 1986). They include multiple sources of evidence with greater weight attached to student and peer input and less weight attached to self-evaluation, alumni, administrators, and others. All of these models are used to arrive at formative and summative decisions.

A Unified Conceptualization

I propose a *unified conceptualization* of teaching effectiveness, whereby evidence is collected from a variety of sources to define the construct and to make decisions about its attainment. Much has been written about the merits and shortcomings of the various sources of evidence currently being employed. Each source can supply unique information, but also is fallible, usually in a way different from the other sources. For example, the unreliability or biases of peer ratings are not the same as those of student ratings; student ratings have other weaknesses. By drawing on three or more different sources of evidence, the strengths of each source can compensate for weaknesses of the other sources, thereby converging on a decision about teaching effectiveness that is more accurate than one based on any single source (Appling, Naumann, & Berk, 2001). This notion of *triangulation* is derived from a compensatory model of decision making. Given the complexity of measuring the act of teaching, it is reasonable to expect that multiple sources can provide a more accurate, reliable, and comprehensive picture of teaching effectiveness than just one source. However, the decision maker should integrate the information from only those sources for which validity evidence is available (see Standard 14.13).

According to Scriven (1991), evaluation is "the process, whose duty is the systematic and objective determination of merit, worth, or value. Without such a process, there is no way to distinguish the worthwhile from the worthless." (p. 4) This process involves two dimensions: (a) gathering data and (b) using that data for judgment and decision making with respect to agreed-upon standards. Measurement tools are needed to collect that data, such as tests, scales, and questionnaires. The criteria for teaching effectiveness are embedded in the content of these measures. The most common measures used for collecting the data for faculty evaluation are rating scales.

12 Sources of Evidence

There are 12 potential sources of evidence of teaching effectiveness: (a) student ratings, (b) peer ratings, (c) self-evaluation, (d) videos, (e) student interviews, (f) alumni ratings, (g) employer ratings, (h) administrator ratings, (i) teaching scholarship, (j) teaching awards, (k) learning outcome measures, and (l) teaching portfolio. An outline of these sources is shown in Table 1 along with several salient characteristics: type of measure needed to gather the evidence, the person(s) responsible for providing the evidence (students, peers, instructor, or administrator), the person or committee who uses the evidence, and the decision(s) typically rendered based on that data (F = formative/ S = summative/ P = program). The purpose of this article is to critically examine the value of these 12 sources reported in the literature on faculty evaluation and to deduce a "bottom line" recommendation for each source based on the current state of research and practice.

Student Ratings

The mere mention of *faculty evaluation* to many college professors conjures up mental images of the "shower scene" from *Psycho*. They're thinking: "Why not just whack me now, rather than wait to see those student ratings again." Student ratings have become synonymous with faculty evaluation in the United States (Seldin, 1999a).

TABLE 1
Salient Characteristics of 12 Sources of Evidence of Teaching Effectiveness

Source of Evidence	Type of Measure(s)	Who Provides Evidence	Who Uses Evidence	Type of Decision[1]
Student Ratings	Rating Scale	Students	Instructors/Administrators	F/S/P
Peer Ratings	Rating Scale	Peers	Instructors	F/S
Self-Evaluation	Rating Scale	Instructors	Instructors/Administrators	F/S
Videos	Rating Scale	Instructors/Peers	Instructors/Peers	F/S
Student Interviews	Questionnaires	Students	Instructors/Administrators	F/S
Alumni Ratings	Rating Scale	Graduates	Instructors/Administrators	F/S/P
Employer Ratings	Rating Scale	Graduates' Employers	Instructors/Administrators	P
Administrator Ratings	Rating Scale	Administrators	Administrators	S
Teaching Scholarship	Judgmental Review	Instructors	Administrators	S
Teaching Awards	Judgmental Review	Instructors	Faculty Committees/Administrators	S
Learning Outcomes	Tests, Projects, Simulations	Students	Instructors/Curriculum Committees	F/P
Teaching Portfolio	Most of the above	Instructors, Students, Peers	Promotions Committees	S

[1]F = formative, S = summative, P = program

It is the most influential measure of performance used in promotion and tenure decisions at institutions that emphasize teaching effectiveness (Emery, Kramer, & Tian, 2003). Recent estimates indicate 88% of all liberal arts colleges use student ratings for summative decisions (Seldin, 1999a). A survey of 40,000 department chairs (US Department of Education, 1991) indicated that 97% used "student evaluations" to assess teaching performance.

This popularity not withstanding, there have also been signs of faculty hostility and cynicism toward student ratings (Franklin & Theall, 1989; Nasser & Fresko, 2002; Schmelkin-Pedhazur, Spencer, & Gellman, 1997). Faculty have lodged numerous complaints about student ratings and their uses. The veracity of these complaints was scrutinized by Braskamp and Ory (1994) and Aleamoni (1999) based on accumulated research evidence. Both reviews found barely a smidgen of research to substantiate any of the common allegations by faculty. Aleamoni's analysis produced a list of 15 "myths" about student ratings. However, there are still dissenters who point to individual studies to support their objections, despite the corpus of evidence to the contrary. At present, a large percentage of faculty in all disciplines exhibit moderately positive attitudes toward the validity of student ratings and their usefulness for improving instruction; however, there's no consensus (Nasser & Fresko, 2002).

There is more research on student ratings than any other topic in higher education (Theall & Franklin, 1990). More than 2000 articles have been cited over the past 60 years (Cashin, 1999; McKeachie & Kaplan, 1996). Although there is still a wide range of opinions on their value, McKeachie (1997) noted that "student ratings are the single most valid source of data on teaching effectiveness" (p. 1219). In fact, there is little evidence of the validity of any other sources of data (Marsh & Roche, 1997). There seems to be agreement among the experts on faculty evaluation that student ratings provides an excellent source of evidence for both formative and summative decisions, with the qualification that other sources also be used for the latter (Arreola, 2000; Braskamp & Ory, 1994; Cashin, 1989, 1990; Centra, 1999; Seldin, 1999a). *[Digression Alert*: If you're itching to be provoked, there are several references on the student ratings debate that may incite you to riot (see Aleamoni, 1999; Cashin, 1999; d'Apollonia & Abrami, 1997; Eiszler, 2002; Emery et al., 2003; Greenwald, 1997; Greenwald & Gilmore, 1997; Greimel-Fuhrmann & Geyer, 2003; Havelka, Neal, & Beasley, 2003; Lewis, 2001; Millea & Grimes, 2002; Read, Rama, & Raghunandan, 2001; Shevlin, Banyard, Davies, & Griffiths, 2000; Sojka, Gupta, & Deeter-Schmelz, 2002; Sproule, 2002; Theall, Abrami, & Mets, 2001; Trinkaus, 2002; Wachtel, 1998). However, before you grab your riot gear, you might want to consider 11 other sources of evidence. *End of Digression].*

BOTTOM LINE: Student ratings is a necessary source of evidence of teaching effectiveness for both formative and summative decisions, but not a sufficient source for the latter. Considering all of the polemics over its value, it is still an essential component of any faculty evaluation system.

Peer Ratings

In the early 1990s, Boyer (1990) and Rice (1991) redefined scholarship to include teaching. After all, it is the means by which discovered, integrated, and applied knowledge is transmitted to the next generation of scholars. Teaching is a scholarly activity. In order to prepare and teach a course, faculty must complete the following:

- Conduct a comprehensive up-to-date review of the literature.
- Develop content outlines.
- Prepare a syllabus.
- Choose the most appropriate print and nonprint resources.
- Write and/or select handouts.
- Integrate instructional technology (IT) support (e.g., audiovisuals, Web site).
- Design learning activities.
- Construct and grade evaluation measures.

Webb and McEnerney (1995) argued that these products and activities can be as creative and scholarly as original research.

If teaching performance is to be recognized and rewarded as scholarship, it should be subjected to the same rigorous peer review process to which a research manuscript is subjected prior to being published in a referred journal. In other words, teaching should be judged by the same high standards applied to other forms of scholarship: *peer review*. Peer review as an alternative source of evidence seems to be climbing up the evaluation ladder, such that more than 40% of liberal arts colleges use peer observation for summative evaluation (Seldin, 1999a).

Peer review of teaching is composed of two activities: peer observation of in-class teaching performance and peer review of the written documents used in a course. *Peer observation of teaching performance* requires a rating scale that covers those aspects of teaching that peers are better qualified to evaluate than students. The scale items typically address the instructor's content knowledge, delivery, teaching methods, learning activities, and the like (see Berk, Naumann, & Appling, 2004). The ratings may be recorded live with one or more peers on one or

multiple occasions or from videotaped classes.

Peer review of teaching materials requires a different type of scale to rate the quality of the course syllabus, instructional plans, texts, reading assignments, handouts, homework, and tests/projects. Sometimes teaching behaviors such as fairness, grading practices, ethics, and professionalism are included. This review is less subjective and more cost-effective, efficient, and reliable than peer observations. However, the observations are the more common choice because they provide direct evaluations of the act of teaching. Both forms of peer review should be included in a comprehensive system, where possible.

Despite the current state of the art of peer review, there is considerable resistance by faculty to its acceptance as a complement to student ratings. Its relative unpopularity stems from the following top 10 reasons:

1. Observations are biased because the ratings are personal and subjective (peer review of research is blind and subjective).
2. Observations are unreliable (peer review of research can also yield low inter-reviewer reliability).
3. One observer is unfair (peer review of research usually has two or three reviewers).
4. In-class observations take too much time (peer review of research can be time-consuming, but distributed at the discretion of the reviewers).
5. One or two class observations does not constitute a representative sample of teaching performance for an entire course.
6. Only students who observe an instructor for 40+ hours over an entire course can really evaluate teaching performance.
7. Available peer rating scales don't measure important characteristics of teaching effectiveness.
8. The results probably will not have any impact on teaching.
9. Teaching is not valued as much as research, especially at large, research-oriented universities; so why bother?
10. Observation data are inappropriate for summative decisions by administrators.

Most of these reasons or perceptions are legitimate based on how different institutions execute a peer review system. A few can be corrected to minimize bias and unfairness and improve the representativeness of observations.

However, there is consensus by experts on reason 10: Peer observation data should be used for formative rather than for summative decisions (Aleamoni, 1982; Arreola, 2000; Centra, 1999; Cohen & McKeachie, 1980; Keig & Waggoner, 1995; Millis & Kaplan, 1995). In fact, 60 years of experience with peer assessment in the military and private industry led to the same conclusion (Muchinsky, 1995). Employees tend to accept peer observations when the results are used for constructive diagnostic feedback instead of as the basis for administrative decisions (Cederblom & Lounsbury, 1980; Love, 1981).

BOTTOM LINE: Peer ratings of teaching performance and materials is the most complementary source of evidence to student ratings. It covers those aspects of teaching that students are not in a position to evaluate. Student and peer ratings, viewed together, furnish a very comprehensive picture of teaching effectiveness for teaching improvement. Peer ratings should not be used for personnel decisions.

Self-Evaluation

How can we ask faculty to evaluate their own teaching? Is it possible for us to be impartial about our own performance? Probably not. It is natural to portray ourselves in the best light possible. Unfortunately, the research on this issue is skimpy and inconclusive. A few studies found that faculty rate themselves higher than (Centra, 1999), equal to (Bo-Linn, Gentry, Lowman, Pratt, and Zhu, 2004; Feldman, 1989), or lower than (Bo-Linn et al., 2004) their students rate them. Highly rated instructors give themselves higher ratings than less highly rated instructors (Doyle & Crichton, 1978; Marsh, Overall, & Kesler, 1979). Superior teachers provide more accurate self-ratings than mediocre or putrid teachers (Centra, 1973; Sorey, 1968).

Despite this possibly biased estimate of our own teaching effectiveness, this evidence can provide support for what we do in the classroom and can present a picture of our teaching unobtainable from any other source. Most administrators agree. Among liberal arts college academic deans, 59% always include self-evaluations for summative decisions (Seldin, 1999a). The Carnegie Foundation for the Advancement of Teaching (1994) found that 82% of four-year colleges and universities reported using self-evaluations to measure teaching performance. The American Association of University Professors (1974) concluded that self-evaluation would improve the faculty review process. Further, it seems reasonable that our assessment of our own teaching should count for something in the teaching effectiveness equation.

So what form should the self-evaluations take? The faculty activity report (a.k.a. "brag sheet") is the most common type of self-evaluation. It describes teaching, scholarship, service, and practice (for the professions) activities for the previous year. This information is used by academic administrators for merit pay decisions. This annual report, however,

is not a true self-evaluation of teaching effectiveness.

When self-evaluation evidence is to be used in conjunction with other sources for personnel decisions, Seldin (1999b) recommends a structured form to display an instructor's teaching objectives, activities, accomplishments, and failures. Guiding questions are suggested in the areas of classroom approach, instructor-student rapport, knowledge of discipline, course organization and planning, and questions about teaching. Wergin (1992) and Braskamp and Ory (1994) offer additional types of evidence that can be collected.

The instructor can also complete the student rating scale from two perspectives: as a direct measure of his or her teaching performance and then as the anticipated ratings the students should give. Discrepancies among the three sources in this triad—students' ratings, instructor's self-ratings, and instructor's perceptions of students' ratings—can provide valuable insights on teaching effectiveness. The results may be very helpful for targeting specific areas for improvement. Students' and self-ratings tend to yield low positive correlations (Braskamp, Caulley, & Costin, 1979; Feldman, 1989).

For formative decisions, the ratings triad may prove fruitful, but a video of one's own teaching performance can be even more informative as a source of self-evaluation evidence. It will be examined in the next section.

Overall, an instructor's self-evaluation demonstrates his or her knowledge about teaching and perceived effectiveness in the classroom (Cranton, 2001). This information should be critically reviewed and compared with the other sources of evidence for personnel decisions. The diagnostic profile should be used to guide teaching improvement.

BOTTOM LINE: Self-evaluation is an important source of evidence to consider in formative and summative decisions. Faculty input on their own teaching completes the triangulation of the three direct observation sources of teaching performance: students, peers, and self.

Videos

Everyone's doing videos. There are cable TV stations devoted exclusively to playing videos. If Britney, Beyonće, and Snoop Dogg can make millions from videos, we should at least make the effort to produce a simple video and we don't have to sing or dance. We simply do what we do best: talk. I mean *teach*.

Find your resident videographer, audiovisual or IT expert, or a colleague who wants to be Steven Spielberg, Ron Howard, or Penny Marshall. Schedule a taping of one typical class or a best and worst class to sample a variety of teaching. Don't perform. Be yourself to provide an authentic picture of how you really teach. The product is a tape or DVD. This is hard evidence of your teaching.

Who should evaluate the video?

1. Self, privately in office, but with access to medications.
2. Self completes peer observation scale of behaviors while viewing, then weeps.
3. One peer completes scale and provides feedback.
4. Two or three peers complete scale on same video and provide feedback.
5. MTV, VH-1, or BET.

These options are listed in order of increasing complexity, intrusiveness, and amount of information produced. All options can provide valuable insights into teaching to guide specific improvements. The choice of option may boil down to what an instructor is willing to do and how much information he or she can handle.

Braskamp and Ory (1994) and Seldin (1999b) argue the virtues of the video for teaching improvement. However, there's only a tad of evidence on its effectiveness. Don't blink or you'll miss it. If the purpose of the video is to diagnose strengths and weaknesses on one or more teaching occasions, faculty should be encouraged to systemically evaluate the behaviors observed using a rating scale or checklist (Seldin, 1998). Behavioral checklists have been developed by Brinko (1993) and Perlberg (1983). They can focus feedback on what needs to be changed. If a skilled peer, respected mentor, or consultant can provide feedback in confidence, that would be even more useful to the instructor (Braskamp & Ory, 1994).

Whatever option is selected, the result of the video should be a profile of positive and negative teaching behaviors followed by a list of specific objectives to address the deficiencies. This direct evidence of teaching effectiveness can be included in an instructor's self-evaluation and teaching portfolio. The video is a powerful documentary of teaching performance.

BOTTOM LINE: If faculty are really committed to improving their teaching, a video is one of the best sources of evidence for formative decisions, interpreted either alone or, preferably, with peer input. If the video is used in confidence for this purpose, faculty should decide whether it should be included in their self-evaluation or portfolio as a "work sample" for summative decisions.

Student Interviews

Group interviews with students furnish another source of evidence that faculty rate as more accurate, trustworthy, useful, comprehensive, and believable than

student ratings and written comments (Braskamp & Ory, 1994), although the information collected from all three sources is highly congruent (Braskamp, Ory, & Pieper, 1980). Faculty consider the interview results as most useful for teaching improvement, but can also be valuable in promotion decisions (Ory & Braskamp, 1981).

There are three types of interviews recommended by Braskamp and Ory (1994): (a) quality control circles, (b) classroom group interviews, and (c) graduate exit interviews. The first type of interview is derived from a management technique used in Japanese industry called *quality control circles* (Shariff, 1999; Weimer, 1990), where groups of employees are given opportunities to participate in company decision making. The instructional version of the "circle" involves assembling a group of volunteer students to meet regularly (bi-weekly) to critique teaching and testing strategies, pinpoint problem areas, and solicit suggestions for improvement. These instructor-led meetings foster accountability for everything that happens in the classroom. The students have significant input into the teaching-learning process and other hyphenated word combos. The instructor can also report the results of the meeting to the entire class to elicit their responses. This opens communication. The unstructured "circle" and class interviews with students on teaching activities can be extremely effective for making changes in instruction. However, faculty must be open to student comments and be willing to make necessary adjustments to improve. This formative evaluation technique permits student feedback and instructional change systematically throughout a course.

Classroom group interviews involves the entire class, but is conducted by someone other than the instructor, usually a colleague in the same department, a graduate TA, or a faculty development or student services professional. The interviewer uses a structured questionnaire to probe the strengths and weaknesses of the course and teaching activities. Some of the questions should allow enough latitude to elicit a wide range of student perspectives from the class. The information collected is shared with the instructor for teaching improvement, but may also be used as a source of evidence for summative decisions.

Graduate exit interviews can be executed either individually or in groups by faculty, administrators, or student services personnel. Given the time needed even for a group interview of undergraduate or graduate students, the questions should focus on information not gathered from the exit rating scale. For example, group interview items should concentrate on most useful courses, least useful courses, best instructors, content gaps, teaching quality, advising quality, and graduation plans. Student responses may be recorded from the interview or may be requested as anonymous written comments on the program. The results should be forwarded to appropriate faculty, curriculum committees, and administrators. Depending on the specificity of the information collected, this evidence may be used for formative feedback and also summative decisions.

BOTTOM LINE: The quality control circle is an excellent technique to provide constant student feedback for teaching improvement. The group interview as an independent evaluation can be very informative to supplement student ratings. Exit interviews may be impractical to conduct or redundant with exit ratings, described in the next section.

Exit and Alumni Ratings

As graduates and alumni, what do students really remember about their instructors' teaching and course experiences? The research indicates: a lot! A longitudinal study by Overall and Marsh (1980) compared "current-student" end-of-term-ratings with one-to-four year "alumni" after-course ratings in 100 courses. The correlation was .83 and median ratings were nearly identical. Feldman (1989) found an average correlation of .69 between current-student and alumni ratings across six cross-sectional studies. This similarity indicates alumni retain a high level of detail about their course taking experiences (Kulik, 2001).

In the field of management, workplace exit surveys and interviews are conducted regularly (Vinson, 1996). Subordinates provide valuable insights on the performance of supervisors. However, in school, exit and alumni ratings of the same faculty and courses will essentially corroborate the ratings given earlier as students. So what should alumni be asked?

E-mailing or snail-mailing a rating scale one, five, and ten years later can provide new information on the quality of teaching, usefulness of course requirements, attainment of program outcomes, effectiveness of admissions procedures, preparation for graduate work, preparation for the real world, and a variety of other topics not measured on the standard student ratings scale. This retrospective evaluation can elicit valuable feedback on teaching methods, course requirements, evaluation techniques, integration of technology, exposure to diversity, and other topics across courses or for the program as a whole. The unstructured responses may highlight specific strengths of faculty as well as furnish directions for improvement. Hamilton, Smith, Heady, and Carson (1995) reported the results of a study of open-ended questions on graduating senior exit surveys. The feedback proved useful to both faculty and administrators. Although this type of survey can tap information beyond "faculty evaluation," such as the curriculum content and sequencing, scheduling of

classes, and facilities, it can be extremely useful as another source of evidence on the quality of teaching on a more generic level.

BOTTOM LINE: Although exit and alumni ratings are similar to original student ratings on the same scale, different scale items about the quality of teaching, courses, curriculum admissions, and other topics can provide new information. Alumni ratings should be considered as another important source of evidence on teaching effectiveness.

Employer Ratings

What "real world" approach to evaluating teaching effectiveness could tap employers' evaluations of graduates? Did they really learn anything from their program of study? Are they successful? After time has passed, at least a year, an assessment (a.k.a. performance appraisal) of the graduate's on-the-job performance can furnish feedback on overall teaching quality, curricular relevance, and program design. Depending on the specificity of the outcomes, inferences may be drawn about individual teaching effectiveness. However, this measure is limited because it is indirect and based on program outcomes.

The first step is to track down the graduates. The admissions office usually maintains records of employment for a few years after graduation. When graduates change jobs or escape to developing countries, PIs and bounty hunters will be needed to find them. Seppanen (1995) suggests using unemployment insurance databases to track graduates' employment history, which can be linked directly to the institution's information systems.

Next, decide what behaviors to measure. Program outcomes can be used when the school is preparing a graduate for a specific profession, such as teaching, nursing, accounting, engineering, football, or espionage. More generic outcomes would be given for the 8,273 other college majors.

These outcomes along with questions about satisfaction with employee performance can be assembled into a rating scale to determine the quality of his or her knowledge, skills, and abilities (KSAs) based on their performance. The ratings across graduates can pinpoint faculty, course, and program strengths and weaknesses in relation to job performance. This can yield mighty useful information.

BOTTOM LINE: Employer ratings provides an indirect source of evidence for program evaluation decisions about teaching effectiveness and attainment of program outcomes, especially for professional schools. Job performance data may be linked to individual teaching performance, but on a very limited basis.

Administrator Ratings

Associate deans, program directors, or department heads can evaluate faculty for annual merit review according to criteria for teaching, scholarship, service, and/or practice (Diamond, 2004). After all, they were or still are faculty with expertise on teaching methods, classroom evaluation techniques, and content in the discipline. The administrator may observe teaching effectiveness and examine documentation in the three other areas, prepared by each faculty member.

Typically, a structured activity report is distributed to all faculty to furnish a comprehensive picture of achievement in all areas over the past year. The more explicit the categories requested in the report, the easier it is for faculty to complete and for administrators to evaluate. The administrators can then rate the overall quality of performance in each category. The total rating across categories can then be scaled to determine merit pay increases.

BOTTOM LINE: Administrator ratings is typically based on secondary sources, not direct observation of teaching or any other areas of performance. This source furnishes a perspective different from all other sources on merit pay and promotion decisions.

Teaching Scholarship

The scholarship of teaching and learning according to the Carnegie Academy for the Scholarship of Teaching an Learning (CASTL), is "a public account of some or all of the full act of teaching—vision, design, enactment, outcomes, and analysis—in a manner susceptible to critical review by the teacher's professional peers and amenable to productive employment in future work by members of the same community" (Shulman, 1998, p. 6). [*Translation*: Contribute to a growing body of knowledge about teaching and learning in higher education by presenting at teaching and learning conferences and publishing in teaching and learning journals.] This scholarship is analogous to scholarship in various disciplines.

Presentations and publications in teaching and learning on innovative teaching techniques and related issues are indicators of teaching expertise. Research on important questions in teaching and learning can not only improve a faculty member's effectiveness in his or her own classroom, but also advance practice beyond it (Hutchings & Shulman, 1999). Evidence of teaching scholarship may consist of presentations on new teaching methods, such as research, workshops, and keynotes, at teaching institutes and conferences. There

are numerous state, regional, national, and international conferences. A few of the best interdisciplinary conferences include the Lilly Conference on College Teaching (plus regional conferences), International Conference on the Scholarship of Teaching and Learning, International Conference on College Teaching and Learning, International Society for Exploring Teaching and Learning Conference, Society for Teaching and Learning in Higher Education Conference (Canadian), and Improving University Teaching Conference. There are also discipline-specific conferences that focus exclusively on teaching and educational issues, such as the National League for Nursing (NLN) Education Summit Conference and Association for Medical Education in Europe (AMEE) Conference.

Publication-wise, there are opportunities to publish in peer-reviewed "teaching" journals. Examples are the *Journal on Excellence in College Teaching*, *College Teaching*, *Journal of Scholarship of Teaching and Learning*, *International Journal of Teaching and Learning in Higher Education, Research in Higher Education*, *Assessment and Evaluation in Higher Education*, and *Creative College Teaching Journal*. There are also more than 50 disciplinary journals (Weimer, 1993).

For faculty who are already conducting research and publishing in their own disciplines, this source of evidence for faculty evaluation provides an opportunity to shift gears and redirect research efforts into the teaching and learning domain. Contributions to scholarship in a discipline *AND* teaching and learning can appreciate a faculty's net worth in two categories rather than in just one.

BOTTOM LINE: Teaching scholarship is an important source of evidence to supplement the three major direct observation sources. It can easily discriminate the "teacher scholar" and very creative faculty from all others for summative decisions.

Teaching Awards

What does this topic have to do with faculty evaluation? That's what I'm here for. Well, the concept is somewhat narrower than the preceding sources of evidence. The link is the process by which the award is determined. A faculty nominee for any award must go through a grueling evaluation by a panel of judges according to criteria for exemplary teaching. The evidence of teaching effectiveness would be limited by the award criteria and review and the pool of nominees.

Estimates in the 1990s indicate that nearly 70% of two-year colleges and liberal arts institutions and 96% of research universities surveyed have awards or programs honoring exemplary teaching (Jenrette & Hayes, 1996; Zahorski, 1996). The literature on the value of teaching awards as an incentive for teaching improvement is sparse (Carusetta, 2001), but runs the gamut from *yes* (Seldin & Associates, 1999; Wright & Associates, 1995) to *no* (McNaught & Anwyl, 1993; Ruedrich, Cavey, Katz, & Grush, 1992; Zahorski, 1996). There has been considerable criticism about the selection process, in particular, which tends to be erratic, vague, suspicious, and subjective (Knapper, 1997; Menges, 1996; Weimer, 1990).

BOTTOM LINE: *As a source of evidence of teaching effectiveness, at best, teaching awards provide worthwhile information only on the nominees, and, at worst, they supply inaccurate and unreliable feedback on questionable nominees who may have appeared on Law and Order. The merits of teaching awards should be evaluated in the context of an institution's network of incentives and rewards for teaching.*

Learning Outcome Measures

Most of the preceding sources of evidence involve direct ratings of teaching behaviors. Learning outcome measures is a sticky source because it is indirect. Teaching performance is being inferred from students' performance—what they learned in the course. Theall and Franklin (2001) noted consistently high correlations between student ratings of "amount learned" and overall ratings. Further, there are significant correlations between student ratings and performance on final exams (Cohen, 1981).

Despite these relationships, establishing student performance on learning outcomes as an independent, valid measure of teaching effectiveness is fraught with numerous difficulties. The crux of the problem is isolating teaching as the sole explanation for student learning. Performance throughout a course on tests, projects, reports, and other indicators may be influenced by the characteristics of the students, the institution, and the outcome measures themselves, over which faculty have no control (Berk, 1988, 1990).

Teaching effectiveness is assessed in terms of student productivity; that is, it is outcomes-based. After all, if a factory worker's performance can be measured by the number of widgets he or she produces over a given period of time, why not evaluate faculty by his or her students' productivity or success on outcome measures? The arguments for this factory worker–teacher productivity analogy are derived from the principles of a piece-rate compensation system (Murnane & Cohen, 1986). Piece-rate contracts is the most common form of "payment by results" (Pencavel, 1977). These contracts provide a strong incentive for workers to produce, because high productivity

results in immediate rewards.

When this system is applied to teaching it breaks down for two reasons. First, a factory worker uses the same materials (e.g., plywood and chewing gum) to make each product (e.g., widget). Faculty work with students whose characteristics vary considerably from class to class. Second, the characteristics of a factory worker's materials rarely influence his or her skills and rate of production; that is, the quality and quantity of widget production can be attributed solely to the worker. Key characteristics of students, such as ability, attitude, motivation, age, gender, and maturation, and of the institution, such as class size, classroom facilities, available technology and learning resources, and school climate, can affect student performance irrespective of what an instructor does in the classroom.

Fenwick (2001) recommends that the results of standard outcome measures, such as tests, problem-solving exercises, projects, and simulations, be aggregated across groups of students for program evaluation decisions about teaching methods and program improvement. Also, multiple measures can be combined to give meaningful feedback to faculty about patterns in outcomes.

BOTTOM LINE: Learning outcome measures should be employed with extreme caution as a source of evidence for faculty evaluation. It's safer to use in conjunction with the direct data sources described previously for program improvement.

Teaching Portfolio

The teaching portfolio is not a single source of evidence; rather, it is a shopping mall of most of the preceding 11 sources assembled systematically for the purpose of promotion and tenure decisions. In fact, portfolio is derived from two Latin root words, "port," meaning "carry," and "folio," meaning "wheelbarrel of best work to the appointments and promotions (A & P) committee with the hope of being promoted." Whew! What a derivation. The term "portfolio" has been associated with the visual arts, architecture, and modeling. It is actually a humongous, skinny, flat, zippered leather case containing photographs, sketches, drawings, securities, and Tyra Banks, which represent an artist's "best work." This package is presented to an editor with the hope of being hired. Huuuum. Are you noting the similarities? Good.

Teaching portfolio is "a coherent set of materials, including work samples and reflective commentary on them, compiled by a faculty member to represent his or her teaching practice as related to student learning and development" (Cerbin & Hutchings, 1993, p. 1). Ahhh. The plot thickens. Now we have two elements to consider: work samples and reflective commentary. If you think this stuff is new and innovative, you're wrong. Work samples have been used in business and industry to measure the performance of employees for more than 50 years. The research on their effectiveness in performance appraisal has been conducted in the field of industrial/organizational psychology (Asher & Sciarrino, 1974; Siegel, 1986). Other definitions contain these basic elements, (Berk, 1999, 2002; Cox, 1995; Edgerton, Hutchings, & Quinlan, 1991; Knapper & Wright, 2001; Murray, 1995; Seldin, Annis, & Zubizarreta, 1995).

Knapper (1995) traced the most recent origins of the teaching portfolio to the work of a committee of the Canadian Association of University Teachers (CAUT). The chair, Shore (1975), argued that faculty should prepare their own evidence for teaching effectiveness – a "portfolio of evidence" (p. 8). What emerged was *The Teaching Dossier: A Guide to Its Preparation and Use* (Shore & Associates, 1980, 1986). In the 1980s, this *Guide* became the portfolio bible and the idea spread like the flu in Canada as the "dossier," in the United States as the "portfolio" (Seldin, 1980, 2004) (*Note*: "dossier" had sinister connotations near the end of Cold War), in Australia (Roe, 1987), and in the United Kingdom as the "profile" (Gibbs, 1988).

So what should we stick in the portfolio-dossier-profile to provide evidence of teaching effectiveness? The *Guide* recommends 49 categories grouped under three headings: (a) Products of good teaching, (b) Material from oneself, and (c) Information from others. Knapper and Wright (2001) offer a list of the 10 most frequently used items from a faculty survey of North American colleges and universities (O'Neil & Wright, 1995):

1. Student course and teaching evaluation data which suggest improvements or produce an overall rating of effectiveness or satisfaction
2. List of course titles and numbers, unit values or credits, enrollments with brief elaboration
3. List of course materials prepared for students
4. Participation in seminars, workshops, and professional meetings intended to improve teaching
5. Statements from colleagues who have observed teaching either as members of a teaching team or as independent observers of a particular course, or who teach other sections of the same course
6. Attempts at instructional innovations and evaluations of their effectiveness
7. Unstructured (and possibly unsolicited) written evaluations by students, including written comments on exams and letters received after a course has been completed
8. Participating in course or curriculum development

9. Evidence of effective supervision on Honors, Master's, or Ph.D. thesis
10. Student essays, creative work, and projects or field work reports (pp. 22–23)

They suggest three categories of items: (a) a statement of teaching responsibilities, (b) a statement of teaching approach or philosophy, and (c) data from students. This is considered a bare bones portfolio.

Before I present my synthesis and bottom line, there is one reaaally important underlying notion that is often overlooked: the portfolio headings and long list of sources of evidence of teaching effectiveness are designed to impress upon the most cynical, imperceptive, biased, and/or ignorant faculty on an A & P committee that *teaching is a scholarly activity* which is comparable to the list of publications, presentations, grants, and research honors presented as evidence of research scholarship. Teaching practice is not just a list of courses and student rating summaries.

Based on a synthesis of components appearing in teaching portfolios cited in the literature and used at several institutions, here is a fairly comprehensive list of elements sorted into three mutually exclusive categories:

1. Description of Teaching Responsibilities
    a. Courses taught
    b. Guest presentations
    c. One-on-one teaching (e.g., scholarly projects, independent studies, thesis/dissertation committees)
    d. Development of new programs or courses
    e. Service on curriculum committees
    f. Training grants
2. Reflective Analysis (5–10 pages)
    a. Philosophy of teaching
    b. Innovative and creative teaching techniques
    c. Mentorship of students and faculty
    d. Participation in faculty development activities
    e. Scholarship of teaching
    f. Recognition of effective teaching
3. Artifacts (Appendices – evidence to support above claims)
    a. Syllabi
    b. Handouts
    c. Exams
    d. Student work samples
    e. Use of technology
    f. Student ratings
    g. Peer ratings
    h. Alumni ratings
    i. Videotapes/DVDs of teaching
    j. Teaching scholarship
    k. Consultations on teaching

Since this portfolio requires considerable time in preparation, its primary use is for career decisions – promotion and tenure (Diamond, 2004; Seldin, 2004). It is a monster self-evaluation compared to the one described previously. Faculty are required to take major responsibility for documenting their teaching accomplishments and practices. Preliminary estimates of the reliability of promotions committee judgments based on portfolios are encouraging (Anderson, 1993; Centra, 1999). The reflective component alone would benefit all faculty if they would take the time to prepare it.

BOTTOM LINE: As a collection of many of the previous sources and them some, the teaching portfolio should be reserved primarily for summative decisions to present a comprehensive picture of teaching effectiveness to complement the list of research publications.

Decision Time

So now that you've surveyed the field of sources, which ones are you going to pick? So many sources, so little time! Which sources already exist in your department? What is the quality of the measures used to provide evidence of teaching effectiveness? Are the faculty stakeholders involved in the current process?

You have some decisions to make. They may be tentative at this point. Use Table 1 and my bottom line recommendations as guides. Transforming the *unified conceptualization* into action means that you

- start with student ratings and one or more other sources that your faculty can embrace which reflect best practices in teaching;
- weigh the pluses and minuses of the different sources (don't bite off too much, but pick as many as possible);
- decide which combination of sources should be used for both formative and summative decisions and those that should be used for one type of decision but not the other, such as peer ratings.

Whatever combination of sources you choose to use, take the time and make the effort to design, execute, and report the results appropriately. The accuracy of faculty evaluation decisions hinges on the integrity of the process and the reliability and validity of the evidence you collect.

References

AERA (American Educational Research Association), APA (American Psychological Association), & NCME (National Council on Measurement in Education) Joint Committee on Standards. (1999).

Standards for educational and psychological testing. Washington, DC: AERA.

Aleamoni, L. M. (1982). Components of the instructional setting. *Instructional Evaluation, 7*, 11–16.

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*, 153–166.

American Association of University Professors. (1974). Committee C. Statement on teaching evaluation. *AAUP Bulletin, 60*(2), 166–170.

Anderson, E. (Ed.). (1993). *Campus use of the teaching portfolio: Twenty-five profiles*. Washington, DC: American Association for Higher Education.

Appling, S. E., Naumann, P. L., & Berk, R. A. (2001). Using a faculty evaluation triad toachieve evidence-based teaching. *Nursing and Health Care Perspectives, 22,* 247–251.

Arreola, R. A. (2000). *Developing a comprehensive faculty evaluation system: A handbook for college faculty and administrators on designing and operating a comprehensive faculty evaluation system* (2nd ed.). Bolton, MA: Anker.

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. *Personnel Psychology, 27,* 519–533.

Berk, R. A. (1988). Fifty reasons why student achievement gain does not mean teacher effectiveness. *Journal of Personnel Evaluation in Education, 1,* 345–363.

Berk, R. A. (1990). Limitations of using student achievement data for career ladder promotions and merit pay decisions. In J. V. Mitchell, Jr., S. L. Wise, & B. S. Plake (Eds.), *Assessment of teaching: Purposes, practices, and implications for the profession* (pp. 261–306). Hillsdale, NJ: Erlbaum.

Berk, R. A. (1999). Assessment for measuring professional performance. In D. P. Ely., L. E. Odenthal, & T. J. Plomp (Eds.), *Educational science and technology: perspectives for the future* (pp. 29–48). Enschede, The Netherlands: Twente University Press.

Berk, R. A. (2002). Teaching portfolios used for high-stakes decisions: You have technical issues! In National Evaluation Systems, *How to find and support tomorrow's teachers* (pp. 45–56). Amherst, MA: Author.

Berk, R. A., Naumann, P. L., & Appling. S. E. (2004). Beyond student ratings: Peer observation of classroom and clinical teaching. *International Journal of Nursing Education Scholarship, 1*(1), 1–26.

Bo-Linn, C., Gentry, J., Lowman, J., Pratt, R. W., & Zhu, R. (2004, November). *Learning from exemplary teachers*. Paper presented at the annual Lilly Conference on College Teaching, Miami University, Oxford, OH.

Boyer, E. (1990). *Scholarship reconsidered: New priorities for the professoriate*. Princeton, NJ: The Carnegie Foundation for the Advancement of Teaching.

Braskamp, L. A., Caulley, D. N., & Costin, F. (1979). Student ratings and instructor self-ratings and their relationship to student achievement. *American Educational Research Journal, 16,* 295–306.

Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work*. San Francisco: Jossey-Bass.

Braskamp, L. A., Ory, J. C., & Pieper, D. M. (1981). Student written comments: Dimensions of instructional quality. *Journal of Educational Psychology, 73*, 65–70.

Brinko, K. T. (1993). The practice of giving feedback to improve teaching: What is effective? *Journal of Higher Education, 64*(5), 54–68.

Carnegie Foundation for the Advancement of Teaching. (1994). *National survey on the reexamination of faculty roles and rewards*. Princeton, NJ: Carnegie Foundation for the Advancement of Teaching.

Carusetta, E. (2001). Evaluating teaching though teaching awards. In C. Knapper & P. Cranton (Eds.), *Fresh approaches to the evaluation of teaching* (New Directions for Teaching and Learning, No. 88) (pp. 31–46). San Francisco: Jossey-Bass.

Cashin, W. E. (1989). *Defining and evaluating college teaching (*IDEA Paper No. 21). Manhattan, KS: Center for Faculty Evaluation and Development, Kansas State University

Cashin, W. E. (1990). *Student ratings of teaching: Recommendations for use* (IDEA Paper No. 22). Manhattan, KS: Center for Faculty Evaluation and Development, Kansas State University.

Cashin, W. E. (1999). Student ratings of teaching: Uses and misuses. In P. Seldin & Associates (Eds.), *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 25–44). Bolton, MA: Anker.

Cederblom, D., & Lounsbury, J. W. (1980). An investigation of user acceptance of peer evaluations. *Personnel Psychology, 33,* 567–580.

Centra, J. A. (1973). Self-ratings of college teachers: A comparison with student ratings. *Journal of Educational Measurement, 10*, 287–295.

Centra, J. A. (1999). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco: Jossey-Bass.

Cerbin, W., & Hutchings, P. (1993, June). *The teaching portfolio*. Paper presented at the Bush Summer Institute, Minneapolis, MN.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51,* 281–309.

Cohen P. A., & McKeachie, W. J. (1980). The role of colleagues in the evaluation of teaching. *Improving College and University Teaching, 28,* 147–154.

Cox, M. D. (1995). A department-based approach to developing teaching portfolios: Perspectives for faculty and development chairs. *Journal on Excellence in College Teaching, 6*(1), 117–143.

Cranton, P. (2001). Interpretive and critical evaluation. In C. Knapper & P. Cranton (Eds.), *Fresh approaches to the evaluation of teaching* (New Directions for Teaching and Learning, No. 88) (pp. 11–18). San Francisco: Jossey-Bass.

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*, 1198–1208.

Diamond, R. M. (2004). *Preparing for promotion, tenure, and annual review: A faculty guide* (2nd ed.). Bolton, MA: Anker.

Doyle, K. O., & Crichton, L. I. (1978). Student, peer, and self-evaluation of college instruction. *Journal of Educational Psychology, 70,* 815–826.

Edgerton, R., Hutchings, P., & Quinlan, K. (1991). *The teaching portfolio: Capturing the scholarship in teaching*. Washington, DC: American Association for Higher Education.

Eiszler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education*, *43*(4), 483–502.

Emery, C. R., Kramer, T. R., & Tian, R. G. (2003). Return to academic standards: A critique of students' evaluations of teaching effectiveness. *Quality Assurance in Education: An International Perspective, 11*(1), 37–47.

Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education, 30,* 137–189.

Fenwick, T. J. (2001). Using student outcomes to evaluate teaching. A cautious exploration. In C. Knapper & P. Cranton (Eds.), *Fresh approaches to the evaluation of teaching* (New Directions for Teaching and Learning, No. 88) (pp. 63–74). San Francisco: Jossey-Bass.

Franklin, J., & Theall, M. (1989, April). *Who read ratings: Knowledge, attitude and practice of users of students' ratings of instruction*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Gibbs, G. (1988). *Creating a teaching profile*. Bristol, England: Teaching and Educational Services.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52,* 1182–1186.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52,*1209–1217.

Greimel-Fuhrmann, B., & Geyer, A. (2003). Students' evaluation of teachers and instructional quality: Analysis of relevant factors based on empirical evaluation research. *Assessment & Evaluation in Higher Education, 28*(3), 229–239.

Hamilton, J. B. III, Smith, M., Heady, R. B., & Carson, P. P. (1997). Using open-ended questions on senior exit surveys to evaluate and improve faculty performance: Results from a school of business administration. *Journal on Excellence in College Teaching, 8*(1), 23–48.

Havelka, D., Neal, C. S., & Beasley, F. (2003, November). *Student evaluation of teaching effectiveness: What criteria are most important?* Paper presented at the annual Lilly Conference College Teaching, Miami University, Oxford, OH.

Hutchings, P., & Shulman, L. S. (1999). The scholarship of teaching: New elaborations, new developments. *Change, 31*(5), 11–15.

Jenrette, M., & Hayes, K. (1996). Honoring exemplary teaching: The two-year college setting. In M. D. Svinicki & R. J. Menges (Eds.), *Honoring exemplary teaching* (New Directions for Teaching and Learning, No. 65). San Francisco: Jossey-Bass.

Keig, L. W., & Waggoner, M. D. (1994). *Collaborative peer review: The role of faculty in improving college teaching* (ASHE/ERIC Higher Education Report, No. 2). Washington, DC: Association for the Study of Higher Education.

Keig, L. W., & Waggoner, M. D. (1995). Peer review of teaching: Improving college instruction through formative assessment. *Journal on Excellence in College Teaching, 6*(3), 51–83.

Knapper, C. (1995). The origins of teaching portfolios. *Journal on Excellence in College Teaching, 6*(1), 45–56.

Knapper, C. (1997). Rewards for teaching. In P. Cranton (Ed.), *University challenges in faculty work: Fresh perspectives from around the world* (New Directions for Teaching and Learning, No. 65). San Francisco: Jossey-Bass.

Knapper, C., & Cranton, P. (Eds.). (2001). *Fresh approaches to the evaluation of teaching* (New Directions for Teaching and Learning, No. 88). San Francisco: Jossey-Bass.

Knapper, C., & Wright, W. A. (2001). Using portfolios to document good teaching: Premises, purposes, practices. In C. Knapper & P. Cranton (Eds.), *Fresh approaches to the evaluation of teaching*

(New Directions for Teaching and Learning, No. 88) (pp. 19–29). San Francisco: Jossey-Bass.

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *The student ratings debate*: *Are they valid*? *How can we best use them*? (New Directions for Institutional Research, No. 109) (pp. 9–25). San Francisco: Jossey-Bass.

Lewis, K. G. (Ed.). (2001). *Techniques and strategies for interpreting student evaluations* (New Directions for Teaching and Learning, No. 87). San Francisco: Jossey-Bass.

Love, K. G. (1981). Comparison of peer assessment methods: Reliability, validity, friendship bias, and user reaction. *Journal of Applied Psychology, 66,* 451–457.

Marsh, H.W., Overall, J. U., & Kesler, S. P. (1979). The validity of students' evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their students. *Journal of Educational Psychology, 71,*149–160.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52,*1187–1197.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52,*1218–1225.

McKeachie, W. J. & Kaplan, M. (1996). Persistent problems in evaluating college teaching. *AAHE Bulletin, 48*(6), 5–8.

McNaught, C., & Anwyl, J. (1993). *Awards for teaching excellence at Australian Universities* (University of Melbourne Centre for the Study of Higher Education Research Working Paper No. 93.1). (ED 368-291)

Menges, R. J. (1996). Awards to individuals. In M. D. Svinicki & R. J. Menges (Eds.), *Honoring exemplary teaching* (New Directions for Teaching and Learning, No. 65). San Francisco: Jossey-Bass.

Millea, M., & Grimes, P. W. (2002). Grade expectations and student evaluation of teaching. *College Student Journal, 36*(4), 582–591.

Millis, B. J., & Kaplan, B. B. (1995). Enhance teaching through peer classroom observations. In P. Seldin & Associates (Eds.), *Improving college teaching* (pp. 137–152). Bolton, MA: Anker.

Muchinsky, P. M. (1995). Peer review of teaching: Lessons learned from military and industrial research on peer assessment. *Journal on Excellence in College Teaching, 6*(3), 17–30.

Murnane, R. J., & Cohen, D. K. (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. *Harvard Educational Review, 56,* 1–17.

Murray, J. P. (1995). The teaching portfolio: A tool for department chairperson to create a climate for teaching excellence. *Innovative Higher Education, 19*(3), 163–175.

Nasser, F., & Fresko, B. (2002). Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education, 27*(2), 187–198.

O'Neil, C., & Wright, W. A. (1995). *Recording teaching accomplishment: A Dalhousie guide to the teaching dossier* (5th ed.). Halifax, Canada: Dalhousie University Office of Instructional Development and Technology.

Ory, J. C., & Braskamp, L. A. (1981). Faculty perceptions of the quality and usefulness of three types of evaluative information. *Research in Higher Education, 15,* 271–282.

Overall, J. U., & Marsh, H. W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology, 72*, 321–325.

Pencavel, J. H. (1997). Work effort, on-the-job screening, and alternative methods of remuneration. In R. Ehrenberg (Ed.), *Research in labor economics* (Vol. 1) (pp. 225–258). Greenwich, CT: JAI Press.

Perlberg, A. E. (1983). When professors confront themselves: Towards a theoretical conceptualization of video self-confrontation in higher education. *Higher Education, 12,* 633–663.

Read, W. J., Rama, D. V., & Raghunandan, K. (2001). The relationship between student evaluations of teaching and faculty evaluations. *Journal of Business Education, 76*(4), 189–193.

Rice, R. E. (1991). The new American scholar: Scholarship and the purposes of the university. *Metropolitan Universities, 1*(4) 7–18.

Romberg, E. (1985). Description of peer evaluation within a comprehensive evaluation program in a dental school. *Instructional Evaluation, 8*(1), 10–16.

Roe, E. (1987). *How to compile a teaching portfolio.*. Kensington, Australia: Federation of Australian University Staff Associations.

Ruedrich, S. L., Cavey, C., Katz, K., & Grush, L. (1992). Recognition of teaching excellence through the use of teaching awards: A faculty perspective. *Academic Psychiatry, 16*(1), 10–13.

Schmelkin-Pedhazur, L., Spencer, K. J., & Gellman, E. S. (1997). Faculty perspectives on course and teacher evaluation. *Research in Higher Education, 38*(5), 575–592.

Scriven, M. (1991). *Evaluation thesaurus* (4th Ed.). Thousand Oaks, CA: Sage.

Seldin, P. (1980). *Successful faculty evaluation programs: A practical guide to improvedfaculty performance and promotion/tenure decisions.* Crugers, NY: Coventry Press.

Seldin, P. (1998, February). *The teaching portfolio.* Paper presented for the American Council on Education, Department Chairs Seminar, San Diego, CA.

Seldin, P. (1999a). Current practices – good and bad – nationally. In P. Seldin & Associates (Eds.), *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 1–24). Bolton, MA: Anker.

Seldin, P. (1999b). Self-evaluation: What works? What doesn't? In P. Seldin & Associates (Eds.), *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 97–115). Bolton, MA: Anker.

Seldin, P. (2004). *The teaching portfolio* (3rd ed.). Bolton, MA: Anker.

Seldin, P., Annis, L., & Zubizarreta, J. (1995). Answers to common questions about the teaching portfolio. *Journal on Excellence in College Teaching, 6*(1), 57–64.

Seldin, P., & Associates (Eds.). (1999). *Changing practices in evaluating teaching: A practical guide to improve faculty performance and promotion/tenure decisions.* Bolton, MA: Anker.

Seppanen, L. J. (1995). Linkages to the world of employment. In P. T. Ewell (Ed.), *Student tracking: New techniques, new demands.* San Francisco: Jossey-Bass.

Shariff, S. H. (1999). Students' quality control circle: A case study on students' participation in the quality of control circles at the Facility of Business and Management. *Assessment & Evaluation in Higher Education, 24*, 141–146.

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: Love me, love my lectures? *Assessment & Evaluation in Higher Education, 25*(4), 397–405.

Shore, B. M. (1975). Moving beyond the course evaluation questionnaire in evaluating university teaching *CAUT Bulletin, 23*(4), 7–10.

Shore, B. M., & Associates. (1980). *The teaching dossier: A guide to its preparation and use.* Ottawa: Canadian Association of University Teachers.

Shore, B. M., & Associates. (1986). *The teaching dossier: A guide to its preparation and use* (rev. ed.). Ottawa: Canadian Association of University Teachers.

Shulman, L. S. (1998). Course anatomy: The dissection and analysis of knowledge through teaching. In P. Hutchings (Ed.), *The course portfolio: How faculty can examine their teaching to advance practice and improve student learning.* Washington, DC: American Association for Higher Education.

Siegel, A. I. (1986). Performance tests. In R. A. Berk (Ed.), *Performance assessment: Methods and applications* (pp. 121–142). Baltimore, MD: Johns Hopkins University Press.

Soderberg, L.O. (1986). A credible model: Evaluating classroom teaching in higher education. *Instructional Evaluation, 8,* 13–27.

Sojka, J., Gupta, A. K., & Deeter-Schmelz, D. R. (2002). Student and faculty perceptions of student evaluations of teaching. *College Teaching, 50*(2), 44–49.

Sorey, K. E. (1968). A study of the distinguishing characteristics of college faculty who are superior in regard to the teaching function. *Dissertation Abstracts, 28*(12-A), 4916.

Sproule, R. (2002). The under-determination of instructor performance by data from the student evaluation of teaching. *Economics of Education Review, 21*(3), 287–295.

Theall, M., Abrami, P. C., & Mets, L. A. (Eds.). (2001). *The student ratings debate: Are they valid? How can we best use them?* (New Directions for Institutional Research, No. 109). San Francisco: Jossey-Bass.

Theall, M., & Franklin, J. L. (1990). Student ratings in the context of complex evaluation systems. In M. Theall & J. L. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice* (New Directions for Teaching and Learning, No. 43). San Francisco: Jossey-Bass.

Theall, M., & Franklin, J. L. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? In M. Theall, P. C., Abrami, & L. A. Mets (Eds.), *The student ratings debate: Are they valid*? *How can we best use them*? (New Directions for Institutional Research, No. 109) (pp. 45–56). San Francisco: Jossey-Bass.

Trinkaus, J., (2002). Students' course and faculty evaluations: An informal look. *Psychological Reports, 91*, 988.

US Department of Education. (1991, Winter). Assessing teaching performance. *The Department Chair: A Newsletter for Academic Administrators, 2*(3), 2.

Vinson, M. N. (1996). The pros and cons of 360-degree feedback: Making it work. *Training and Development, 50*(4), 11–12.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 23*, 199–212.

Webb, J., & McEnerney, K. (1995). The view from the back of the classroom: A faculty-based peer observation program. *Journal on Excellence in College Teaching, 6*(3), 145–160.

Weimer, M. E. (1990). *Improving college teaching: Strategies for developing instructional effectiveness*. San Francisco: Jossey-Bass.

Weimer, M. E. (1993). The disciplinary journal of pedagogy. *Change, 25,* 44–51.

Wergin, J. E. (1992, September). *Developing and using performance criteria*. Paper presented at the Virginia Commonwealth University Conference on Faculty Rewards, Richmond.

Wright, A. W., & Associates (1995). *Teaching improvement practices: Successful strategies for higher education*. Bolton, MA: Anker.

Zahorski, K. J. (1996). Honoring exemplary teaching in the liberal arts institution. In M. D. Svinicki & R. J. Menges (Eds.), *Honoring exemplary teaching*. (New Directions for Teaching and Learning, No. 65). San Francisco: Jossey-Bass.

RONALD A. BERK, PhD, is Professor of Biostatistics and Measurement at the School of Nursing, The Johns Hopkins University. He served as Assistant Dean for Teaching from 1997-2003. He received the University's Alumni Association Excellence in Teaching Award in 1993 and Caroline Pennington Award for Teaching Excellence in 1997 and was inducted as a Fellow in the Oxford Society of Scholars in 1998. He has published 8 books, including 2 on humor: *Professors Are from Mars*, *Students Are from Snickers* (Stylus, 2003) and *Humor as an Instructional Defibrillator* (Stylus, 2002). The quality of those books and his more than 120 journal/book publications and 200 presentations reflects his life-long commitment to mediocrity and his professional motto: "Go for the Bronze!"