5. Identifying Exemplary Teachers and Teaching: Evidence from Student Ratings[1]

Kenneth A. Feldman
SUNY-Stony Brook University
Kenneth.Feldman.1@sunysb.edu

Key Words: College teaching; dimensions of teaching; exemplary teaching; student ratings of instruction; reliability and validity; teaching myths vs. research evidence; faculty evaluation; faculty development

Formal or systematic evaluation by college students of their teachers has long been used to help students in their selection of courses, to provide feedback to faculty about their teaching, and to supply information for administrators and personnel committees in their deliberations on the promotion and tenure of individual faculty members. Moreover, with the increasing emphasis that many colleges and universities are currently putting on good teaching and on designating, honoring, and rewarding good teachers, the use of student ratings is, if anything, likely to increase. Yet, for all their use, student ratings of instructors and instruction are hardly universally accepted. It is no secret, for example, that some college teachers have little regard for them. For these faculty, student evaluations of teachers (or courses)—whether sponsored by the university administration, faculty-development institutes, individual academic departments, or student-run organizations—are not reliable, valid, or useful, and may even be

[1]This paper is based on an earlier one (Feldman, 1994) commissioned by the National Center on Post-secondary Teaching, Learning, and Assessment for presentation at the Second AAHE Conference on Faculty Roles & Rewards held in New Orleans (January 28–30, 1994). The earlier paper benefited by the thoughtful suggestions of Robert Menges and Maryellen Weimer. As for the present paper, I am grateful to Herbert Marsh, Harry Murray, and Raymond Perry for their helpful comments. A brief version of this paper is to appear in an issue of *New Directions for Teaching and Learning*, edited by Marill Svinicki and Robert Menges (Feldman, forthcoming).

R.P. Perry and J.C. Smart (eds.), The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective, 93–143.

harmful. Others, of course, believe more or less the opposite; and still others fall somewhere in between these two poles of opinion.

If the credibility of teacher evaluations is to be based on more than mere opinion, one asks what the research on their use shows. This question turns out to be more difficult to answer than might be thought because, even apart from the substance of the pertinent research, the number of relevant studies is voluminous. A few years ago, in a letter to the editor in *The Chronicle of Higher Education* (Sept. 5, 1990), William Cashin pointed out that 1,300 citations could be found in the Educational Resources Information Center on "student evaluation of teacher performance" at the postsecondary level. This same year, my own collection of books and articles on instructional evaluation numbered about 2,000 items (Feldman, 1990b). This collection has grown still larger since then, of course. It is true that, at a guess, well over one-half of the items in this collection are opinion pieces (filled with insightful observations at best and uninformed polemics at worst). Even so, this still leaves a large number of research pieces.

Luckily, this research—either as a whole or subportions of it—has been reviewed relatively often (see, among others, Aubrect, 1981; Braskamp, Brandenburg and Ory, 1984; Braskamp and Ory, 1994; Centra, 1979, 1989, 1993; Costin, Greenough and Menges, 1971; Doyle, 1975, 1983; Kulik and McKeachie, 1975; Marsh, 1984, 1987; Marsh and Dunkin, 1992; McKeachie, 1979, Miller, 1972, 1974; and Murray, 1980). Cashin (1988, 1995) has even supplied particularly useful reviews of the major reviews. My own series of reviews started in the mid-1970s and has continued to the present. (See Feldman, 1976a, 1976b, 1977, 1978, 1979, 1983, 1984, 1986, 1987, 1989a, 1989b, 1990a, 1993; two other analyses—Feldman, 1988, 1992—are indirectly relevant.)

One of the best overviews in the area is that by Marsh (1987), which is an update and elaboration of an earlier review of his (Marsh, 1984). In this review, after 100 pages or so of careful, critical, and reflective analysis of the existing research and major reviews of student ratings of instruction, Marsh (1987) sums up his findings and observations, as follows:

> Research described in this article demonstrates that student ratings are clearly multidimensional, quite reliable, reasonably valid, relatively uncontaminated by many variables often seen as sources of potential bias, and are seen to be useful by students, faculty, and administrators. However, the same findings also demonstrate that student ratings may have some halo effect, have at least some unreliability, have only modest agreement with some criteria

> of effective teaching, are probably affected by some potential sources of bias and are viewed with some skepticism by faculty as a basis for personnel decisions. It should be noted that this level of uncertainty probably also exists in every area of applied psychology and for all personnel evaluation systems. Nevertheless, the reported results clearly demonstrate that a considerable amount of useful information can be obtained from student ratings; useful for feedback to faculty, useful for personnel decision, useful to students in the selection of courses, and useful for the study of teaching. Probably, students' evaluations of teaching effectiveness are the most thoroughly studied of all forms of personnel evaluation, and one of the best in terms of being supported by empirical research (p. 369).

Marsh's tempered conclusions set the stage for the present comments. This discussion first explores various interpretations that can be made of information gathered from students about their teachers (which includes a consideration of the possible half-truths and myths that continue to circulate about teacher and course evaluations). It then analyzes the differential importance of the individual items that constitute the rating forms used to evaluate teachers. The primary aim of this discussion is to see how student evaluations can be used to help identify exemplary teachers and instruction.

TRUTHS, HALF-TRUTHS, AND MYTHS: INTERPRETING STUDENT RATINGS

The unease felt by some faculty, and perhaps by some administrators and students as well, in using teacher and course evaluations to help identify exemplary teachers and instruction may in part be due to the half-truths if not outright myths that have cropped up about these evaluations. Some of the myths can be laid to rest; and the half-truths can be more fully analyzed to separate the real from the imagined. To do so requires a consideration of certain factors or influences that have been said to "bias" ratings. At the moment there is no clear consensus on the definition of bias in the area of student ratings (see Marsh, 1984, 1987; and Marsh and Dunkin, 1992). I take bias to mean something other than (or more than) the fact that student ratings may be influenced by conditions not under the teacher's control or that conditions may somehow be "unfair" to the instructor (making it harder for him or her to teach well and thus to get high ratings compared to teachers in "easier" situations). Rather,

bias here refers to one or more factors directly and somehow inappropriately influencing students' judgments about and evaluation of teachers or courses. In essence, the question is whether a condition or influence actually affects teachers and their instruction, which is then accurately reflected in students' evaluations (a case of *non*bias), or whether in some way this condition or influence only affects students' attitudes toward the course and students' perceptions of instructors (and their teaching) such that evaluations do not accurately reflect the instruction that students receive (a case of bias). (For a more extensive discussion of the meaning of bias as it pertains to student ratings, see Feldman, 1984, 1993; Marsh, 1987, and Marsh and Dunkin, 1992.) Implications and examples of this conceptualization of bias will be given as the discussion proceeds.

Myths

Aleamoni (1987) has listed a number of speculations, propositions, and generalizations about students' ratings of instructors and instruction that he declares "are (on the whole) myths." Although I would not go so far as to call each of the generalizations on his list a myth, some of them indeed are—at least as far as current research shows—as follows: students cannot make consistent judgments about the instructor and instruction because of their immaturity, lack of experience, and capriciousness (untrue); only colleagues with excellent publication records and expertise are qualified to teach and to evaluate their peers' instruction—good instruction and good research being so closely allied that it is unnecessary to evaluate them separately (untrue); most student rating schemes are nothing more than a popularity contest, with the warm, friendly, humorous instructor emerging as the winner every time (untrue); students are not able to make accurate judgments until they have been away from the course, and possibly away from the university for several years (untrue); student ratings are both unreliable and invalid (untrue); the time and day the course is offered affect student ratings (untrue); students cannot meaningfully be used to improve instruction (untrue). I call these statements untrue because supporting evidence was not found for them in one or another of the following research reviews: Abrami, Leventhal, and Perry (1982); Cohen (1980b); Feldman (1977, 1978, 1987, 1989a, 1989b); Levinson-Rose and Menges (1981); L'Hommedieu, Menges and Brinko (1988, 1990); Marsh (1984, 1987); and Marsh and Dunkin (1992).

For the most part, Aleamoni (1987) also seems correct in calling the following statement a myth: "Gender of the student and the

instructor affects student ratings." Consistent evidence cannot be found that either male or female college students routinely give higher ratings to teachers (Feldman, 1977). As for the gender of the teacher, a recent review (Feldman, 1993) of three dozen or so studies showed that a majority of these studies found male and female college teachers not to differ in the global ratings they receive from their students. In those studies in which statistically significant differences were found, more of them favored women than men. However, across all studies, the average association between gender and overall evaluation of the teacher, while favoring women, is so small (average $r = +.02$) as to be insignificant in practical terms. This would seem to show that the gender of the teacher does not bias students' ratings (unless, of course, it can be shown by *other* indicators of teachers' effectiveness that the ratings of one gender "should" be higher than the other to indicate the reality of this group's better teaching).

This said, it should also be noted that there is some indication of an interaction effect between the gender of the student and the gender of the teacher: across studies, there is some evidence to suggest that students may rate same-gendered teachers a little more highly they than do opposite-gendered teachers. What is unknown from the existing studies, however, is what part of this tendency is due to male and female students taking different classes (and thus having different teachers) and what part is due to differences in preferences of male and female students within classes (thus possibly indicating a bias in their ratings).

HALF-TRUTHS AND THE QUESTION OF BIAS IN RATINGS

Aleamoni (1987) also presents the following statements as candidates for the status of myth: the size of the class affects student ratings; the level of the course affects student ratings; the rank of the instructor affects student ratings; whether students take the course as a requirement or as an elective affects their ratings; whether students are majors or nonmajors affects their ratings. That these are myths is not clear-cut. Each of these course, instructor or student factors is, in fact, related to student evaluation. The real question is: "Why?"

Although the results of pertinent studies are somewhat mixed, some weak trends can be discerned: *slightly* higher ratings are given (a) to teachers of smaller rather than larger courses (Feldman, 1984; Marsh, 1987); (b) to teachers of upper-level rather than lower-level courses (Feldman, 1978); (c) to teachers of higher rather than lower academic ranks (Feldman, 1983; Marsh, 1987); (d) by students taking a course

as an elective rather than as a requirement (Feldman, 1978; Marsh, 1987); and (e) by students taking a course that is in their major rather than one that is not (Feldman, 1978; Marsh, 1987). These associations do not prove causation, of course; each of these factors may not actually and directly "affect" ratings, but may simply be associated with the ratings due to their association with other factors affecting ratings.

Even if it can be shown that one or more of these factors actually and directly "affect" students' ratings, the ratings are not necessarily biased by these factors, as is often inferred when such associations are found (probably an important underlying worry of those prone to discount teacher or course evaluations). To give an example, at certain colleges and universities teachers of higher rank may in fact typically be somewhat better teachers, and thus "deserve" the slightly higher ratings they receive. To give another example, teachers in large classes may receive slightly lower ratings because they indeed are somewhat less effective in larger classes than they are in smaller classes, not because students take out their dislike of large classes by rating them a little lower than they otherwise would. So, while it may be somewhat "unfair" to compare teachers in classes of widely different sizes, the unfairness lies in the difference in teaching conditions, not in a rating bias as defined here.[1]

To put the matter in general terms, certain course characteristics and situational contexts—conditions that may not necessarily be under full control of the teachers—may indeed affect teaching effectiveness; and student ratings may then accurately reflect differences in teaching effectiveness. Although rating bias may not necessarily be involved, those interested in using teaching evaluations to help in decisions about promotions and teaching awards may well want to take into account the fact that it may be somewhat harder to be effective in some courses than in others. Along these lines, note that student ratings gathered from the Instructional Development and Effectiveness Assessment (IDEA) system are reported *separately* for four categories of class size—small (1–14 students), medium (15–34), large (35–99) and very large (100 or more)—as well as for five levels of student motivation for the class as a whole (determined by the average of the students' responses to the background question, "I have a strong desire to take this course"). The reason for this procedure is made clear to users of the evaluation instrument, as follows:

[1] Using a different definition of bias, Cashin (1988) would consider the size of class a source of bias if its correlation with student ratings of teachers were sufficiently large (but see Cashin, 1995).

> In addition to using flexible criteria, the IDEA system also controls for *level of student motivation* or the students' desire to take the course. . .and the *size of the class*—two variables which the research has shown are correlated with student rating. . . .The IDEA system assumes that it is harder to teach large groups of students who do not want to take a course than it is to teach small groups of students who do want to take a course. IDEA controls for this by comparing an instructor's ratings, not only with "All" courses in the comparative data pool, but with "Similar" courses [same level of student motivation and same class size] as well (Cashin and Sixbury, 1993, pp. 1–2, emphasis in original).

Another candidate for the status of myth concerns students' grades. As Aleamoni (1987) words it, "the grades or marks students receive in the course are highly correlated with their ratings of the course and instructor." On the one hand, the word "highly" indeed makes the statement mythical; grades are not *highly* correlated with students' ratings. On the other hand, almost all of the available research does show a small or even modest positive association between grades and evaluation (usually a correlation somewhere between +.10 and +.30), whether the unit of analysis is the individual student or the class itself (see Feldman, 1976a, 1977; Stumpf and Freedman, 1979).

Research has shown that some part of the positive correlation between students' grades (usually expected grades) and students' evaluation of teachers is due to "legitimate" reasons and therefore is unbiased: students who learn more earn higher grades and thus legitimately give higher evaluations. This has been called the "validity hypothesis" or "validity effect" (see Marsh, 1987, and Marsh and Dunkin, 1992). Moreover, some part of the association may be spurious, attributable to some third factor—for example, students' interest in the subject matter of the course—which has been referred to as the "student characteristics hypothesis" or "student characteristics effect" (see Marsh, 1989, and Marsh and Dunkin, 1992). Yet another part of the positive correlation may indeed be due to a rater bias in the ratings, although the bias might not be large. Researchers currently are trying to determine the degree to which an attributional bias (students' tendency to take credit for successes and avoid blame for failure) and a retributional bias (students "rewarding" teachers who give them higher grades by giving them higher evaluations, and "punishing" teachers who give them lower grades by giving them lower evaluations) are at work (see Gigliotti and Buchtel, 1990; Theall, Franklin, and Ludlow, 1990a, 1990b). The second of these two biases has been called a

"grading leniency hypothesis" or "grading leniency effect" (Marsh, 1987; Marsh and Dunkin, 1992). In their review of research relating grades and teacher evaluations, Marsh and Dunkin (1992) conclude as follows:

> Evidence from a variety of different types of research clearly supports the validity hypothesis and the student characteristics hypothesis, but does not rule out the possibility that a grading leniency effect operates simultaneously. Support for the grading leniency effect was found with some experimental studies, but these effects were typically weak and inconsistent, may not generalize to nonexperimental settings where SETs [students' evaluations of teaching effectiveness] are actually used, and in some instances may be due to the violation of grade expectations that students had falsely been led to expect or that were applied to other students in the same course. Consequently, while it is possible that a grading leniency effect may produce some bias in SETs, support for this suggestion is weak and the size of such an effect is likely to be insubstantial in the actual use of SETs (p. 202).

Yet another correlate of—and, therefore, a possible influence on—teacher evaluations is not mentioned by Aleamoni (1987): academic discipline of the course. Reviewing eleven studies available at the time (Feldman, 1978), I found that teachers in different academic fields tend to be rated somewhat differently. Teachers in English, humanities, arts, and language courses tend to receive somewhat higher student ratings than those in social science courses (especially political sciences, sociology, psychology and economic courses); this latter group of teachers in turn receive somewhat higher ratings than teachers in the sciences (excepting certain subareas of biological sciences), mathematics and engineering courses. Recently, based on data from tens of thousands of classes either from the IDEA system only (Cashin and Clegg, 1987; Cashin and Sixbury, 1993) or from this system and the Student Instructional Report (SIR) of the Educational Testing Service combined (Cashin, 1990), differences among major fields similar to those in my review have been reported.

Cashin and his associates have suggested several possible causes that could be operating to produce these differences in ratings of teachers in different academic disciplines, including the following: some courses are harder to teach than others; some fields have better teachers than others; and students in different major fields rate differently because of possible differences in their attitudes, academic skills, goals, motivation, learning styles, and perceptions of the

constituents of good teaching. The following practical advice given by Cashin and Sixbury (1993) is informative:

> There is increasing evidence that different academic fields are rated differently. What is not clear is why. Each institution should examine its own data to determine to what extent the differences found in the general research hold true at that particular institution. If an institution concludes that the differences found at that institution are *due to something other than the teaching effectiveness of the instructors*, e.g., because low rated courses are more difficult to teach, or reflect a stricter rating response set on the part of the students taking those courses, then some control for those differences should be instituted. Using the comparative data in this technical report is one possibility. If however, it is decided that the *differences in ratings primarily reflect differences in teaching effectiveness*, that is, that the low rated courses are so rated because they are *not* as well taught, then of course no adjustments should be made (pp. 2–3, emphases in original).

IDENTIFYING INSTRUCTIONAL DIMENSIONS IMPORTANT TO EFFECTIVE TEACHING

Thus far, I have explored how student ratings can be used to identify those persons who are seen by students as exemplary teachers (as well as those who are not), noting certain precautions in doing so. Now, I turn to the related topic of how exemplary teaching itself can be identified through the use of student ratings of specific pedagogical dispositions, behaviors and practices of teachers.[2] Teaching comprises many different elements—a multidimensionality that instruments of teacher evaluation usually attempt to capture. The construction of most of these instruments, as Marsh and Dunkin (1992) point out, is based on "a logical analysis of the content of effective teaching and the purposes the ratings are intended to serve, supplemented by reviews of previous research and feedback" (p. 146). Less often used is an

[2] As with overall evaluation of teachers, the characteristics of courses, of teachers themselves, and of situational contexts have all been found to correlate with specific evaluations. Those characteristics most frequently studied have been class size, teacher rank/experience and the gender of the teacher. Class size and the rank/experience of the teacher each correlate more highly with some specific evaluations than with others (for details, see Feldman, 1983, 1984). (The degree to which these factors actually affect teaching rather than "biasing" students in their ratings has yet to be determined.) With the possible exception of their sensitivity to and concern with class level and progress, male and female teachers do not consistently differ in the specific evaluations they receive across studies (Feldman, 1993).

empirical approach that emphasizes statistical techniques such as factor analysis or multitrait-multimethod analysis.

Marsh and Dunkin (1992) also note that "for feedback to teachers, for use in student course selection, and for use in research in teaching . . . there appears to be general agreement that a profile of distinct components of SETs [students' evaluations of teaching effectiveness] based on an appropriately constructed multidimensional instrument is more useful than a single summary score" (p. 146). However, whether a multidimensional profile score is more useful than a single summary score for personnel decisions has turned out to be more controversial (see Abrami, 1985, 1988, 1989a, 1989b; Abrami and d'Apollonia, 1991; Abrami, d'Apollonia, and Rosenfield, 1993, 1996; Cashin and Downey, 1992; Cashin, Downey, and Sixbury, 1994; Hativa and Raviv, 1993; and Marsh, 1987, 1991a, 1991b, 1994).

In earlier reviews (Feldman, 1976b, 1983, 1984, 1987, 1989a), I used a set of roughly 20 instructional dimensions into which the teaching components of relevant studies could be categorized. In recent years, I extended this set in one way or another to include more dimensions (see Feldman, 1988, 1989b, 1993). The fullest set—28 dimensions—is given in the Appendix, along with specific examples of evaluation items that would be categorized in each dimension. Unlike studies using factor analyses or similar techniques to arrive at instructional dimensions, the categories are based on a logical analysis of the single items and multiple-item scales found in the research literature on students' views of effective teaching and on their evaluations of actual teachers. Over the years, I have found the system of categorization to be useful in classifying the characteristics of instruction analyzed in various empirical studies even though it may differ from the definitions and categories found in any one of these studies.[3]

Teaching That is Associated with Student Learning

Although all 28 dimensions of instruction found in the Appendix would seem to be important to effective teaching, one would assume that some of them are more important than others. One way of establishing this differential importance is to see how various teaching

[3] Abrami and d'Apollonia (1990) adapted these categories for use in their own work (also see d'Apollonia and Abrami, 1988). More recently, they have made more extensive refinements and modifications to the dimensions and concomitant coding scheme (Abrami, d'Apollonia, and Rosenfield, 1993, 1996).

dimensions relate to student learning, which Cohen (1980a, 1981, 1987) did in his well-known meta-analytic study of the relationships of student achievement with eight different instructional dimensions.[4] Based in large part on work by d'Apollonia and Abrami (1987, 1988) and Abrami, Cohen, and d'Apollonia, 1988), I extended Cohen's meta-analysis a few years ago by using less heterogeneous categories for coding the evaluation items and scales in the studies under review, widening the range of instructional dimensions under consideration, and preserving more of the information in the studies Cohen used in his meta-analysis (see Feldman, 1989b, 1990a). To be included in Cohen's meta-analysis or my own, a study had to provide data from actual college classes rather than from experimental analogues of teaching. The unit of analysis in the study had to be the class or instructor and not the individual student. Its data had to be based on a multisection course with a common achievement measure used for all sections of the course (usually an end-of the course examination as it turned out). Finally, the study had to provide data from which a rating/achievement correlation could be calculated (if one was not given).

The correlations between specific evaluations and student achievement from the studies under review were distributed among 28 instructional dimensions (given in the present Appendix), with weighting procedures used to take into account evaluational items or scales that were coded in more than one dimension. Average correlations were calculated for each of the instructional dimensions having information from at least three studies. These average correlations are given in Table 1,T1 along with the percent of variance explained (r^2).[5]

Note that average r's for the instructional dimensions range from +.57 to −.11. All but one (Dimension No. 11) are positive, and all but three (Dimensions No. 11, No. 23, No. 24) are statistically significant. The two highest correlations of .57 and .56—explained variance of over 30%—are for Dimensions No. 5 (teacher's preparation and course organization) and No. 6 (teacher's clarity and understandableness). The teacher's pursuit and/or meeting of course objectives and the student-perceived outcome or impact of the course (Dimensions No. 28 and No. 12) are the next most highly related dimensions with achievement ($r = +.49$ and $+.46$). Somewhat more moderately-sized correlations—indicating between roughly 10% and 15% of explained

[4] These dimensions are labeled: Skill; Rapport; Structure; Difficulty; Interaction; Feedback; Evaluation; and Interest/Motivation.

[5] The results given in Table 1 are similar to those shown in an analysis in d'Apollonia and Abrami (1988), although there are some differences (see Abrami, d'Apollonia, and Rosenfield, 1996).

Table 1: Average Correlations of Specific Evaluations of Teachers with Student Achievement

Percent Variance Explained		Instructional Dimension	Average r
30.0%-34.9%	No. 5	Teacher's Preparation; Organization of the Course	.57
	No. 6	Clarity and Understandableness	.56
25.0%-29.9%			
20.0%-24.9%	No. 28	Teacher Pursued and/or Met Course Objectives	.49
	No. 12	Perceived Outcome or Impact of Instruction	.46
15.0%-19.9%			
10.0%-14.9%	No. 1	Teacher's Stimulation of Interest in the Course and Its Subject Matter	.38
	No. 20	Teacher Motivates Students to Do Their Best; High Standard of Performance Required	.38
	No. 16	Teacher's Encouragement of Questions, and Openness to Opinions of Others	.36
	No. 19	Teacher's Availability and Helpfulness	.36
	No. 7	Teacher's Elocutionary Skills	.35
	No. 9	Clarity of Course Objectives and Requirements	.35
	No. 3	Teacher's Knowledge of the Subject	.34
5.0%-9.9%	No.8	Teacher's Sensitivity to, and Concern with, Class Level and Progress	.30
	No. 2	Teacher's Enthusiasm (for Subject or for Teaching)	.27
	No. 13	Teacher's Fairness; Impartiality of Evaluation of Students; Quality of Examinations	.26
	No. 25	Classroom Management	.26
	No. 17	Intellectual Challenge and Encouragement of Independent Thought (by the Teacher and the Course)	.25
	No. 14	Personality Characteristics ("Personality") of the Teacher	.24
	No. 18	Teacher's Concern and Respect for Students; Friendliness of the Teacher	.23
	No. 15	Nature, Quality, and Frequency of Feedback from the Teacher to the Students	.23
	No. 26	Pleasantness of Classroom Atmosphere	.23

Table 1: *(Continued)*

0.0%-4.9%	No. 10	Nature and Value of the Course (Including Its Usefulness and Relevance)	.17
	No. 23	Difficulty of the Course (and Workload)—Description	.09
	No. 24	Difficulty of the Course (and Workload)—Evaluation	.07
	No. 11	Nature and Usefulness of Supplementary Materials and Teaching Aids	−.11

Note: This table has been constructed from data given in Table 1 in Feldman (1989b), which itself was based on information in the following studies: Benton and Scott (1976); Bolton and Marr (1979); Braskamp, Caulley, and Costin (1979); Bryson (1974); Centra (1977); Chase and Keene (1979); Cohen and Berger (1970); Costin (1978); Doyle and Crichton (1978); Doyle and Whitely (1974); Elliott (1950); Ellis and Rickard (1977); Endo and Della-Piana (1976); Frey (1973); Frey (1976); Frey, Leonard, and Beatty (1975); Greenwood, Hazelton, Smith, and Ware (1976); Grush and Costin (1975); Hoffman (L978); Marsh, Fleiner, and Thomas (1975); Marsh and Overall (1980); McKeachie, Lin and Mann (1971); Mintzes (1976–77); Morgan and Vasché (1978); Morsh, Burgess, and Smith (1956); Murray (1983); Orpen (1980); Rankin, Greenmun, and Tracy (1965); Remmers, Martin, and Elliott (1949); Rubinstein and Mitchell (1970); Solomon, Rosenberg, and Bezdek (1964); and Turner and Thompson (1974). Each r given in (or derived from information in) individual studies was converted to a Fisher's Z transformation (z_r) and weighted by the inverse of the number of instructional dimensions in which it was coded. For each instructional dimension, the weighted z_r's were averaged and then backtransformed to produce the weighted average r's given in this table. These r's are shown only for those instructional dimensions having information from at least three studies; thus there are no entries for Dimensions 4, 21, 22 and 27. All correlations in this table are statistically significant except those for Dimensions 11, 23, and 24.

variance—were found for several instructional dimensions: teacher's stimulation of students' interest in the course and its subject (Instructional Dimension No. 1, average $r = +.38$); teacher's motivation of students to do their best (No. 20, +.38); teacher's encouragement of questions and discussion, and openness to the opinions of others (No. 16, +.36); teacher's availability and helpfulness (No. 19, +.36); teacher's elocutionary skills (No. 7, +.35); clarity of course objectives and requirements (No. 9, +.35); and teacher's knowledge of subject (No. 3, +.34).

Less strongly associated with student achievement are: the teacher's sensitivity to, and concern with, class level and progress (No. 8); teacher's enthusiasm (No. 2); teacher's fairness and impartiality

of evaluation (No. 13); classroom management (No. 25); intellectual challenge and encouragement of students' independent thought (No. 17); teacher's "personality" (No. 14); teacher's friendliness and respect or concern for students (No. 18); the quality and frequency of teacher's feedback to students (No. 15); the pleasantness of the classroom atmosphere (No. 26); and the nature and value of the course material (No. 10). The nature and usefulness of supplementary materials and teaching aids as well as the difficulty and workload of the course (either as a description or as an evaluation by students) are not related to student achievement. Because of insufficient data in the set of studies under consideration, the relationship of the following dimensions to student achievement is not clear from these studies: No. 4 (teacher's intellectual expansiveness); No. 21 (teacher's encouragement of self-initiated learning); No. 22 (teacher's productivity in research); and No. 27 (individualization of teaching).

Do Certain Kinds of Teaching Actually Produce Student Achievement?

It is important to recognize that the associations between specific evaluations of teachers and student achievement by themselves do not establish the causal connections between the instructional characteristics under investigation and student achievement. For example, it is possible that the correlations that have been found in some proportion of the studies (whose results were used to create Table 1) do not necessarily indicate that the instructional characteristics were causal in producing the students' achievement. Rather, as Leventhal (1975) was one of the first to point out, some third variable such as student motivation, ability or aptitude of the class might independently affect both teacher performance and student learning, which would account for the correlations between instructional characteristics and student achievement even if there were no direct causal connection.

Leventhal (1975) has suggested that causality can be more clearly established in studies in which students are randomly assigned to sections of a multisection course rather than self-selected into them, for the "random assignment of students. . .promotes equivalence of the groups of students by disrupting the causal processes which ordinarily control student assignment" (p. 272). It is not always possible, however, to assign students randomly to class sections. In some of the studies reviewed by Cohen (and by Feldman, 1989b), students were randomly assigned to class sections, whereas in other studies they were

not. Interestingly, in his meta-analysis, Cohen (1980a) found that, for each of the four instructional dimensions that he checked, studies in which students were randomly assigned to sections gave about the same results as did studies where students picked their own class sections. Cohen (1980a) also compared studies where the ability of students in class sections was statistically controlled with studies where it was not. Again, for each of the four instructional dimensions that he checked, the correlations for the two sets of studies did not differ. Results such as these increase the likelihood that the instructional characteristics and student achievement are causally connected, although the possibility of spurious elements has not been altogether ruled out. Even with random assignment, the results of multisection validation studies may still permit certain elements of ambiguity in interpretation and generalization (Marsh, 1987; and Marsh and Dunkin, 1992; but see Abrami, d'Apollonia, and Rosenfield, 1993, 1996).

The results of experimental studies—whether field experiments or laboratory experiments–are obviously useful here, for they can help clarify cause-effect relationships in ways that the correlational studies just reviewed cannot. Relevant research has been reviewed (selectively) by Murray (1991), who notes in his analysis of pertinent studies that either teacher's enthusiasm/expressiveness or teacher clarity (or both) has been a concern in nearly all relevant experimental research, and that these studies usually include measures of amount learned by students. In his overview of this research, Murray (1991) reports that "classroom teaching behaviors, at least in the enthusiasm and clarity domains, appear to be *causal antecedents* (rather than mere correlates) of various instructional outcome measures" (p. 161, emphasis added).

Although Murray's (1991) definitions of these domains are not completely identical with the definitions of pertinent dimensions of the present analysis, it is still of interest to compare his conclusions and the findings given here. Thus, in the present discussion, teacher clarity has also been shown to be of high importance to teaching, whether indicated by the correlation of teacher clarity with student achievement in the multisection correlational studies or, as will be seen in a later section of this paper, by the association of teacher clarity with the global evaluation of the teacher. As for the enthusiastic/expressive attitudes and behaviors of teachers, highlighted in Murray's (1991) analysis, the instructional dimensions of "teachers enthusiasm (for subject or for teaching)" referred to in the present discussion is, in fact, associated with achievement in the multisection correlational studies, but only moderately so compared to some of the other

instructional dimensions. However, the instructional dimension of "teacher's elocutionary skills," which assumedly is an aspect of enthusiasm/expressiveness is more strongly associated with achievement in the multisectional-correlational studies. Furthermore, note that Murray writes that "behaviors loading on the Enthusiasm [Expressive] factor share elements of spontaneity and stimulus variation, and thus are perhaps best interpreted as serving to elicit and maintain student attention to material presented in class" (p. 146). Given this interpretation, it is of relevance that the instructional dimension of "teacher's stimulation of interest in the course and its subject matter" has been found to be rather highly correlated (albeit less so than the top four dimensions) with students' achievement in multisectional correlational studies; moreover, this particular dimension is highly associated, as well, with global evaluation of instruction relative to the other instructional dimensions (to be discussed in a later section of this paper).

Underlying Mechanisms and Other Considerations

Whether the associations between student learning and teacher's attitudes, behaviors, and practices are established by correlational studies or by experimental studies, the exact psychological and social psychological mechanisms by which these instructional characteristics influence student learning need to be more fully and systematically detailed than they have been. When a large association between an instructional characteristic and student achievement is found, the tendency is to see the finding as obvious—that is, as being a self-explanatory result. For example, given the size of the correlation involved, it would seem obvious that a teacher who is clear and understandable naturally facilitates students' achievement; little more needs to be said or explained, it might be thought. But, in a very real sense, the "obviousness" or "naturalness" of the connection appears only after the fact (of a substantial association). Were the correlation between dimension of "feedback" and student achievement a great deal larger than was found, then this instructional characteristic, too, would be seen by some as obviously facilitative of student achievement: naturally, teachers who give frequent and good feedback effect high cognitive achievement in their students. But, as previously noted, frequency and quality of feedback has not been found to correlate particularly highly with student achievement, and there is nothing natural or obvious about either a high or low association between feedback and students'

achievement; and, in fact, to see either as natural or obvious ignores the specific psychological and social psychological mechanisms that may be involved in either a high or low correlation.

In short, although a case can be made that many of the different instructional characteristics could be expected to facilitate student learning (see, for example, Marsh and Dunkin, 1992, pp. 154–156), what is needed are specific articulations about which particular dimensions of instruction theoretically and empirically are more likely and which less likely to produce achievement. A crucial aspect of this interest is specifying exactly how those dimensions that affect achievement do so—even when, at first glance, the mechanisms involved would seem to be obvious. Indeed, conceptually and empirically specifying such mechanisms in perhaps the most "obvious" connection of them all in this area—that between student achievement and the clarity and understandableness of instructors—has turned out to be particularly complex, not at all simple or obvious (see, for example, Land, 1979, 1981; Land and Combs, 1981, 1982, Land and Smith, 1979, 1981; and Smith and Land, 1980). Likewise, the mechanisms underlying the correlation between teacher's organization and student achievement have yet to be specifically and fully determined, although Perry (1991) has recently started the attempt by offering the following hypothetical linkages:

> Instructor organization. . .involves teaching activities intended to structure course material into units more readily accessible from students' long-term memory. An outline for the lecture provides encoding schemata and advanced organizers which enable students to incorporate new, incoming material into existing structures. Presenting linkages between content topics serves to increase the cognitive integration of the new material and to make it more meaningful, both of which should facilitate retrieval. (p. 26)

One other consideration may be mentioned at this point. McKeachie (1987) has recently reminded educational researchers and practitioners that the achievement tests assessing student learning in the sorts of studies being considered here typically measure lower-level educational objectives such as memory of facts and definitions rather than the higher-level outcomes such as critical thinking and problem solving that are usually taken as important in higher education. He points out that "today cognitive and instructional psychologists are placing more and more emphasis upon the importance of the way in which knowledge is structured as well as upon skills and strategies

for learning and problem solving" (p. 345). Moreover, although not a consideration of this paper, there are still other cognitive skills and intellectual dispositions as well as a variety of affective and behavioral outcomes of students that my be influenced in the college classroom (see, for example, discussions in Baxter Magolda, 1992; Bowen, 1977; Chickering and Reisser, 1993; Doyle, 1972; Ellner and Barnes, 1983; Feldman and Newcomb, 1969; Feldman and Paulsen, 1994; Hoyt, 1973; King and Kitchener, 1994; Marsh, 1987; Pascarella and Terenzini, 1991; Sanders and Wiseman, 1990; Sockloff, 1973; and Turner, 1970).

Specific Aspects of Teaching as Related to Overall Evaluation of the Teacher

There is another way of determining the differential importance of various instructional dimensions, one that uses information internal to the evaluation form itself. If it is assumed that each student's overall evaluation of an instructor is an additive combination of the student's evaluation of specific aspects of the teacher and his or her instruction, weighted by the student's estimation of the relative importance of these aspects to good teaching, then it would be expected that students' overall assessment of teachers would be more highly associated with instructional characteristics that students generally consider to be more important to good teaching than with those they consider to be less important (cf. Crittenden and Norr, 1973). Thus, one way to establish the differential importance of various instructional characteristics is to compare the magnitudes of the correlations between the actual overall evaluations by students of their teachers and their ratings of each of the specific attitudinal and behavioral characteristics of these teachers. Otherwise put, the importance of an instructional dimension is indicated by its ability to discriminate among students' global assessment of teachers.[6]

In an analysis (Feldman, 1976b) done a while ago now, though one still of full relevance here, I located some 23 studies containing correlations (or comparable information showing the extent of the associations) between students' overall evaluations of their teachers and their ratings of specific attitudinal and behavioral characteristics of these teachers.

[6] Limitations of this approach to determining the importance of instructional dimensions are discussed in Feldman (1976b, 1988; also see Abrami, d'Apollonia and Rosenfield, 1993, 1996).

This information in each study was used to rank order the importance of these characteristics (in terms of size of its association with overall evaluation) and then to calculate for each study standardized ranks (rank of each item divided by the number of items ranked) for the specific evaluations in the study. Finally, for each of the instructional dimensions under consideration (see Feldman, 1976, Table 1 and note 5), standardized ranks were averaged across the pertinent studies.

These average standardized ranks are given in Column 2 of Table 2.T2 Column 1 of this same table repeats those data previously given in Table 1 on the associations between instructional dimensions and student achievement for just those instructional dimensions considered in both analyses. The two analyses, each determining the importance of instructional dimensions in its own way, have eighteen instructional dimensions in common, although data for only seventeen of them are given in the table. Instructional Dimension No. 4 (teacher's intellectual expansiveness) has been left out, as it was in Table 1, because of insufficient data about the correlation between it and student achievement. Table 2T2 also shows (in parentheses) the rank in importance of each of the instructional dimensions that is produced by each of the two different methods of gauging importance of the dimensions.

There is no overlap in the studies on which the data in Columns 1 and 2 of Table 2 are based. Furthermore, because the studies considered in the student achievement analyses (Col. 1) are mostly of students in multisection courses of an introductory nature, these students and courses are less representative of college students and courses in general than are the students and courses in the second set of studies (Col. 2). Despite these circumstances, the rank-order correlation (rho) between the ranks shown in the two columns is +.61. Those specific instructional dimensions that are the most highly associated with student achievement tend to be the same ones that best discriminate among teachers with respect to the overall evaluation they receive from students. The correlation is not a perfect one, however. The largest discrepancies are for teacher's availability and helpfulness (relatively high importance in terms of its association with achievement and relatively low importance in terms of its association student's global evaluations) and for intellectual challenge and encouragement of students' independent thought (relatively low importance by the first indicator and relatively high importance by the second indicator). The other large "shifts" between the two indicators of importance are less dramatic: teacher's preparation and course organization (from Rank 1

Table 2: Comparison of Instructional Dimensions on Two Different Indicators of Importance

	Instructional Dimension	Importance Shown by Correlation with Student Achievement (1)	Importance Shown by Correlation with Overall Evaluations (2)
No. 5	Teacher's Preparation; Organization of the Course	.57 (1)	.41 (6)
No. 6	Clarity and Understandableness	.56 (2)	.25 (2)
No. 12	Perceived Outcome or Impact of Instruction	.46 (3)	.28 (3)
No. 1	Teacher's Stimulation of Interest in the Course and Its Subject Matter	.38 (4)	.20 (1)
No. 16	Teacher's Encouragement of Questions and Discussion, and Openness to Opinions of Others	.36 (5.5)	.60 (11)
No. 19	Teacher's Availability and Helpfulness	.36 (5.5)	.74 (16)
No. 7	Teacher's Elocutionary Skills	.35 (7.5)	.49 (10)
No. 9	Clarity of Course Objectives and Requirements	.35 (7.5)	.45 (7)
No. 3	Teacher's Knowledge of the Subject	.34 (9)	.48 (9)
No. 8	Teacher's Sensitivity to, and Concern with, Class Level and Progress	.30 (10)	.40 (5)
No. 2	Teacher's Enthusiasm (for Subject or for Teaching)	.27 (11)	.46 (8)
No. 13	Teacher's Fairness; Impartiality of Evaluation of Students; Quality of Examinations	.26 (12)	.72 (14.5)
No. 17	Intellectual Challenge and Encouragement of Independent Thought (by the Teacher and the Course)	.25 (13)	.33 (4)
No. 18	Teacher's Concern and Respect for Students; Friendliness of the Teacher	.23 (14.5)	.65 (12)
No. 15	Nature, Quality, and Frequency of Feedback from the Teacher to Students	.23 (14.5)	.87 (17)

Table 2: *(Continued)*

	Instructional Dimension	Importance Shown by Correlation with Student Achievement (1)	Importance Shown by Correlation with Overall Evaluations (2)
No. 10	Nature and Value of the Course Material (Including Its Usefulness and Relevance)	.17 (16)	.70 (13)
No. 11	Nature and Usefulness of Supplementary Materials and Teaching Aids	−.11 (17)	.72 (14.5)

Note: This table is adapted from Table 3 in Feldman (1989b). The correlations shown in Column 1 are the same as those in Table 1 of the present analysis. The higher the correlation, the more important the instructional dimension. The correlations have been ranked from 1 to 17 (with the ranks shown in parentheses). The average standardized ranks given in Column 2 originally were given in Feldman (1976b, see Table 2 and footnote 5), and are based on information in the following studies: Brooks, Tarver, Kelley, Liberty, and Dickerson (1971); Centra (1975); Cobb (1956); French-Lazovik (1974, two studies); Garber (1964); Good (1971); Harry and Goldner (1972); Harvey and Barker (1970); Jioubu and Pollis (1974); Leftwich and Remmers (1962); Maas and Owen (1973); Owen (1967); Plant and Sawrey (1970); Remmers (1929); Remmers and Weisbrodt (1964); Rosenshine, Cohen, and Furst (1973); Sagen (1974); Spencer (1967); Van Horn (1968); Walker (1968); Widlak, McDaniel, and Feldhusen (1973); and Williams (1965). The lower the average standardized rank (that is, the smaller the fraction), the more important the dimension. The average standardized ranks in Column 2 have been ranked from 1 to 17 (with the ranks shown in parentheses). This table includes only those dimensions considered in both Feldman (1976b) and Feldman (1989b), and thus there are fewer dimensions in this table than there are in Table 1.

to Rank 6, the latter still relatively high in importance), and teacher's encouragement of questions and openness to others' opinion (from rank 5.5 to rank 11).

If ranks 1 through 6 are thought of as indicating high importance (relative to the other dimensions), rank 7–12 as indicating moderate importance, and ranks 13–17 as indicating low importance (low, that is, relative to the other dimensions, not necessarily unimportant), then the two methods determining the importance of instructional dimensions show the following pattern. Both methods indicate that the teacher's preparation and course organization, the teacher's clarity and

understandableness, the teacher's stimulation of students' interest and the students' perceived outcome or impact of the course are of high importance (relative to the other dimensions). Although the teacher's encouragement of questions and openness to others' opinion as well as his or her availability and helpfulness are also of high importance in terms of the association of each with achievement, the first is only of moderate importance and the second of low importance in terms of its association with global evaluation of teachers.

Both methods of determining the importance of the instructional dimensions show the following to be of moderate importance relative to other dimensions: teacher's elocutionary skill, clarity of course objective and requirements, teacher's knowledge of subject, and teacher's enthusiasm. The importance of the teacher's sensitivity to class level and progress is also moderate by the first indicator (association with student learning) but high by the second (association with overall evaluation of the teacher), whereas the teacher's fairness and impartiality of evaluation is moderate by the first and low by the second. Each of the following five dimensions is of low relative importance in terms of its association with student achievement, although only the first three are also relatively low in importance in terms of their association with global evaluation: nature, quality and frequency of feedback to students; nature and value of course material; nature and usefulness of supplementary materials and teaching aids; intellectual challenge and encouragement of independent thought (which is of relatively high importance in the strength of its association with the global evaluation of teachers); and teacher's friendliness and concern/respect for student (of moderate importance in its association with global evaluation).

Table 3T3 offers a summary of the results of using the two different ways considered here of determining the importance of various instructional dimensions from student ratings of teachers. By averaging (when possible) the rank order of the dimensions produced by the two methods, information in Table 2 (and, in some cases, Table 1 as well) has been used to classify roughly the instructional dimensions into four categories of importance: high importance; moderate importance; moderate-to-low importance; and low (or no) importance. For most of the instructional dimensions, placement into the categories depended on information from both indicators of importance (association with achievement and association with global rating); in the other cases, classification was based on information from only one indicator (association with achievement).

Table 3: Summary of the Importance of Various Instructional Dimensions Based on Student Ratings

		High Importance
(Two Sources)	No. 6	Clarity and Understandableness
(Two Sources)	No. 1	Teacher's Stimulation of Interest in the Course and Its Perceived Subject Matter
(Two Sources)	No. 12	Perceived Outcome of Impact of Instruction
(Two Sources)	No. 5	Teacher's Preparation; Organization of the Course
(One Source)	No. 28	Teacher Pursued and/or Met Course Objectives
(One Source)	No. 20	Teacher Motivates Students to Do Their Best; High Standard of Performance Required
		Moderate Importance
(Two Sources)	No. 9	Clarity of Course Objectives and Requirements
(Two Sources)	No. 8	Teacher's Sensitivity to, and Concern with, Class Level and Progress
(Two Sources)	No. 16	Teacher's Encouragement of Questions and Discussion, and Openness to Opinions of Others
(Two Sources)	No. 17	Intellectual Challenge and Encouragement of Independent Thought
(Two Sources)	No. 7	Teacher's Elocutionary Skills
(Two Sources)	No. 3	Teacher's Knowledge of the Subject
(Two Sources)	No. 2	Teacher's Enthusiasm for the Subject
(Two Sources)	No. 19	Teacher's Availability and Helpfulness
		Moderate-to-Low Importance
(Two Sources)	No. 13	Teacher's Fairness; Impartiality of Evaluation of Students; Quality of Examinations
(Two Sources)	No. 18	Teacher's Concern and Respect for Students; Friendliness of the Teacher
(One Source)	No. 25	Classroom Management
(One Source)	No. 14	Personality Characteristics ("Personality") of the Teacher
(One Source)	No. 26	Pleasantness of Classroom Atmosphere
		Low Importance or No Importance
(Two Sources)	No. 10	Nature and Value of the Course (Including its Usefulness and Relevance)
(Two Sources)	No. 15	Nature, Quality, and Frequency of Feedback from the Teacher to the Student

(*cont.*)

Table 3: *(Continued)*

(Two Sources)	No. 11	Nature and Usefulness of Supplementary Materials and Teaching Aids
(One Source)	No. 23	Difficulty of the Course (and Workload)—Description
(One Source)	No. 24	Difficulty of the Course (and Workload)—Evaluation

Note: By averaging (when possible) the rank ordering of dimensions produced by two different methods of determining importance of various instructional dimensions, information in Table 2 (and, in some cases, Table 1) has been used to classify instructional dimensions into one of the four categories shown in this table. As indicated in the table, for some instructional dimensions two sources of information were available (association of the instructional dimension with achievement and with global evaluations, as given in Table 2); for other instructional dimensions, only one source of information was available (association of the instructional dimension with achievement, as given in Table 1.)

Although the present paper has concentrated on data derived from student ratings of actual teachers, I want to note briefly another way of determining the importance of various instructional dimensions using different information: Those most involved with teaching and learning can be asked directly about the importance of various components of instruction. In one analysis (Feldman, 1988), I collected thirty-one studies in which both students and faculty (separately) specified the instructional characteristics they considered particularly important to good teaching and effective instruction. Students and faculty were generally similar, though not identical, in their views, as indicated by an average correlation of +.71 between them in their valuation of various aspects of teaching. However, the ordering of the instructional dimensions by either of these groups shows differences (as well as some similarities) with that based on the two indicators of importance using student ratings of actual teachers.

A few examples may be given. Similar to the results shown in Table 3, Instructional Dimensions No. 5 (teacher's preparation and organization of the course) and No. 6 (clarity and understandableness) are of high importance to students and to faculty when these groups are asked directly about what is important to good teaching and effective instruction. Further, when asked directly, students again place high importance on Dimension No. 1 (teacher's stimulation of interest), but in this case faculty (when asked directly) see this aspect of teaching as less important than do the students (when asked directly) or by the two indicators of importance derived from student evaluations (summarized in Table 3). Moreover, compared to the importance determined

by the analysis of data from student evaluations, students and faculty, when asked directly, place less importance on Instructional Dimension No. 12 (perceived outcome or impact of instruction) but more importance on Dimensions No. 8 (teacher's sensitivity to, and concern with, class level and progress), No. 3 (teacher's knowledge of subject matter), and No. 2 (teacher's enthusiasm).[7]

CONCLUDING COMMENTS

This paper was not intended as a comprehensive review of the research literature on evaluation of college students of their teachers or on the correlates of effective teaching in college. Indeed, several topics or areas usually explored in such reviews have not been considered in this paper. To take two instances, I have ignored an analysis of whether there is a connection between research productivity and teaching effectiveness as well as a discussion of the usefulness of student ratings as feedback to faculty to improve their teaching (other than to label as myths the statements that good instruction and good research are so closely allied as to make it unnecessary to evaluate them separately and that student ratings cannot meaningfully be used to improve teaching). Rather, I have somewhat single-mindedly focused on the use of student ratings to identify exemplary teachers and teaching. In doing so, I have drawn together relevant parts of my own work over the years in addition to incorporating findings and conclusions from selected others.

Nothing I have written in this paper is meant to imply that the use of teacher evaluations is the only means of identifying exemplary teachers and teaching at the college level. The recent discussion of the multitude of items that would be appropriate for "teaching portfolios" by itself suggests otherwise (see, among others, Centra, 1993, Edgerton, Hutchings and Quinlan, 1991, and Seldin, 1991). For instance, in a project sponsored by the Canadian Association of University Teachers to identify the kinds of information a faculty member might use as evidence of teaching effectiveness, some forty-nine specific items were suggested as possible items for inclusion in a dossier (Shore and associates, 1986); only one of these items

[7] Other similarities and differences can be found in Feldman, 1989b (Table 3), where data for all four indicators of the importance of various instructional dimensions—association with achievement, association with global ratings, direct report of students, and direct report of faculty—are given.

referred to student ratings (listed as "student course and teaching evaluation data. . ."). Given the diverse ways noted in these dossiers of "capturing the scholarship of teaching," as Edgerton, Hutchings and Quinlan (1991) put it, gathering teacher evaluations may or may not be the one best way to identify excellence in teaching. But it is an important way; and current research evidence does show that when teacher evaluation forms are properly constructed and administered (Feldman, 1979), the global and specific ratings contained in them, as interpreted with appropriate caution, are undeniably helpful in identifying exemplary teachers and teaching.

Reprinted by permission of Agathon Press, New York.

REFERENCES

Abrami, P.C. (1985). Dimensions of effective college instruction. *Review of Higher Education* 8: 211–228.

Abrami, P.C. (1988), SEEQ and ye shall find: A review of Marsh's "Students' evaluation of university teaching." *Instructional Evaluation* 9: 19–27.

Abrami, P.C. (1989a). How should we use student ratings to evaluate teaching? *Research in Higher Education* 30: 221–227.

Abrami, P.C. (1989b). SEEQing the truth about student ratings of instruction. *Educational Researcher* 43: 43–45.

Abrami, P.C., Cohen, P.A., and d'Apollonia, S. (1988). Implementation problems in meta-analysis. *Review of Educational Research* 58: 151–179.

Abrami, P.C., and d'Apollonia, S. (1990). 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.

Abrami, P.C., and d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness—generalizability of "N = 1" research: Comments on Marsh. *Journal of Educational Psychology* 83: 411–415.

Abrami, P.C., d'Apollonia, S., and Rosenfield, S. (1993). *The Dimensionality of Student Ratings of Instruction: Introductory Remarks*. Paper presented at the annual meeting of the American Educational Research Association.

Abrami, P.C., d'Apollonia, S., and Rosenfield, S. (1996). The dimensionality of student ratings of instruction: What we know and what we do not. In J.C. Smart (ed.) *Higher Education: Handbook of Theory and Research* (Vol. 11). New York: Agathon Press.

Abrami, P.C., Leventhal, L., and Perry R.P. (1982). Educational seduction. *Review of Educational Research* 52: 446–464.

Aleamoni, L. (1987). Student rating myths versus research facts. *Journal of Personnel Evaluation in Education* 1: 111–119.

Aubrecht, J.D. (1981). Reliability, validity and generalizability of student ratings of instruction. (IDEA Paper No. 6). Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development. (ERIC Document Reproduction Service No. ED 213 296)

Baxter Magolda, M.B. (1992). *Knowing and Reasoning in College: Gender-Related Patterns in Students' Intellectual Development*. San Francisco: Jossey-Bass.

Benton, S.E., and Scott, O. (1976). *A Comparison of the Criterion Validity of Two Types of Student Response Inventories for Appraising Instruction*. Paper presented at the annual meeting of the National Council on Measurement in Education.

Bolton, B., Bonge, D., and Marr, J. (1979). Ratings of instruction, examination performance, and subsequent enrollment in psychology courses. *Teaching of Psychology* 6: 82–85.

Bowen, H.R. (1977). *Investment in Learning: The Individual and Social Value of American Higher Education*. San Francisco: Jossey-Bass.

Braskamp, L.A., Brandenburg, D.C., and Ory, J.C. (1984). *Evaluating Teaching Effectiveness: A Practical Guide*. Beverly Hills, Calif.: Sage.

Braskamp, L.A., Caulley, D., and 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., and Ory, J.C. (1994). *Assessing Faculty Work: Enhancing Individual and Institutional Performance*. San Francisco: Jossey-Bass.

Brooks, T.E., Tarver, D.A., Kelley, H.P., Liberty, P.G., Jr., and Dickerson, A.D. (1971). Dimensions underlying student ratings of courses and instructors at the University of Texas at Austin: Instructor Evaluation Form 2. (Research Bulletin RB-71-4). Austin, Texas: University of Texas at Austin, Measurement and Evaluation Center.

Bryson, R. (1974). Teacher evaluations and student learning: A reexamination. *Journal of Educational Research* 68: 11–14.

Cashin, W.E. (1988). *Student Ratings of Teaching: A Summary of the Research*. (IDEA Paper No. 20). Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W.E. (1990). Students do rate different academic fields differently. 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.

Cashin, W.E. (1995). *Student Ratings of Teaching: The Research Revisited*. (IDEA Paper No. 32). Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W.E., and Clegg, V.L. (1987). *Are Student Ratings of Different Academic Fields Different?* Paper presented at the annual meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 286 935)

Cashin, W.E., and Downey, R.G. (1992). Using global student rating items for summative evaluation. *Journal of Educational Psychology* 84: 563–572.

Cashin, W.E., Downey, R.G., and Sixbury, G.R. (1994). Global and specific ratings of teaching effectiveness and their relation to course objectives: Reply to Marsh. *Journal of Educational Psychology* 86: 649–657.

Cashin, W.E., and Sixbury, G.R. (1993). *Comparative Data by Academic Field*. (IDEA Technical Report No. 8). Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

Centra, J.A. (1975). Colleagues as raters of classroom instruction. *Journal of Higher Education* 46: 327–337.

Centra, J.A. (1977). Student ratings of instruction and their relationship to student learning. *American Educational Research Journal* 14: 17–24.

Centra, J.A. (1979). *Determining Faculty Effectiveness: Assessing Teaching, Research, and Service for Personnel Decisions and Improvement*. San Francisco: Jossey-Bass.

Centra, J.A. (1989). Faculty evaluation and faculty development in higher education. In J.C. Smart (ed.), *Higher Education: Handbook of Theory and Research* (Vol. 5). New York: Agathon Press.

Centra, J.A. (1993). *Reflective Faculty Evaluation: Enhancing Teaching and Determining Faculty Effectiveness*. San Francisco: Jossey-Bass.

Chase, C.I., and Keene, J.M., Jr. (1979). Validity of student ratings of faculty. (Indiana Studies in Higher Education No. 40). Bloomington, Ind.: Indiana University, Bureau of Evaluation Studies and Testing, Division of Research and Development. (ERIC Document Reproduction Service No. ED 169 870).

Chickering, A.W., and Reisser, L. (1993). *Education and Identity* (2nd edition). San Francisco: Jossey-Bass.

Cobb, E.B. (1956). *Construction of a Forced-choice University Instructor Rating Scale*. Unpublished doctoral dissertation, University of Tennessee, Knoxville.

Cohen, P.A. (1980a). *A Meta-analysis of the Relationship between Student Ratings of Instruction and Student Achievement*. Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

Cohen, P.A. (1980b). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education* 13: 321–341.

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

Cohen, P.A. (1987). *A Critical Analysis and Reanalysis of the Multisection Validity Meta-analysis*. Paper presented at the annual meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 283 876)

Cohen, S.H., and Berger, W.G. (1970). Dimensions of students' ratings of college instructors underlying subsequent achievement on course examinations. *Proceedings of the 78th Annual Convention of the American Psychological Association* 5: 605–606.

Costin, F. (1978). Do student ratings of college teachers predict student achievement? *Teaching of Psychology* 5: 86–88.

Costin, F., Greenough, W.T., and Menges, R. J. (1971). Student ratings of college teaching: Reliability, validity and usefulness. *Review of Educational Research* 41: 511–535.

Crittenden, K.S., and Norr, J.L. (1973). Student values and teacher evaluation: A problem in person perception. *Sociometry* 36: 143–151.

d'Apollonia, S., and Abrami, P.C. (1987). *An Empirical Critique of Metaanalysis: The Literature on Student Ratings of Instruction*. Paper presented at the annual meeting of the American Educational Research Association.

d'Apollonia, S., and Abrami, P.C. (1988). *The Literature on Student Ratings of Instruction: Yet Another Meta-analysis*. Paper presented at the annual meeting of the American Educational Research Association.

d'Apollonia, S., Abrami, P.C., and Rosenfield, S. (1993). *The Dimensionality of Student Ratings of Instruction: A Meta-Analysis of the Factor Studies*. Paper presented at the annual meeting of the American Educational Research Association.

Doyle, K.O., Jr. (1972). *Construction and Evaluation of Scale for Rating College Instructors*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Doyle, K.O., Jr. (1975). *Student Evaluation of Instruction*. Lexington, MA: D. C. Heath.

Doyle, K.O., Jr. (1983). *Evaluating Teaching*. Lexington, MA: D. C. Health.

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

Doyle, K.O., Jr., and Whitely, S.E. (1974). Student ratings as criteria for effective teaching. *American Educational Research Journal* 11: 259–274.

Edgerton, R., Hutchings, P., and Quinlan, K. (1991). *The Teaching Portfolio: Capturing the Scholarship in Teaching*. Washington, DC: American Association for Higher Education.

Elliott, D.N. (1950). Characteristics and relationship of various criteria of college and university teaching. *Purdue University Studies in Higher Education* 70: 5–61.

Ellis, N.R., and Rickard, H.C. (1977). Evaluating the teaching of introductory psychology. *Teaching of Psychology* 4: 128–132.

Ellner, C.L., and Barnes, C.P. (1983). *Studies of College Teaching: Experimental Results, Theoretical Interpretations, and New Perspectives*. Lexington, MA: D. C. Heath.

Endo, G.T., and Della-Piana, G. (1976). A validation study of course evaluation ratings. *Improving College and University Teaching* 24: 84–86.

Feldman, K.A. (1976a). Grades and college students' evaluation of their courses and teachers. *Research in Higher Education* 4: 69–111.

Feldman, K.A. (1976b). The superior college teacher from the students' view. *Research in Higher Education* 5: 243–288.

Feldman, K.A. (1977). Consistency and variability among college students in rating their teachers and courses: A review and analysis. *Research in Higher Education* 6: 223–274.

Feldman, K.A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education* 9: 199–242.

Feldman, K.A. (1979). The significance of circumstances for college students' ratings of their teachers and courses. *Research in Higher Education* 10: 149–172.

Feldman, K.A. (1983). Seniority and experience of college teachers as related to evaluation they receive from students. *Research in Higher Education* 18: 3–124.

Feldman, K.A. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education* 21: 45–116.

Feldman, K.A. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: A review and synthesis. *Research in Higher Education* 24: 139–213.

Feldman, K.A. (1987). Research productivity and scholarly accomplishment of college teachers as related to their instructional effectiveness: A review and exploration. *Research in Higher Education* 26: 227–298.

Feldman, K.A. (1988). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities? *Research In Higher Education* 28: 291–344.

Feldman, K.A. (1989a). 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: 137–194.

Feldman, K.A. (1989b). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education* 30: 583–645.

Feldman, K.A. (1990a). An afterword for "The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies." *Research in Higher Education* 31: 315–318.

Feldman, K.A. (1990b). Instructional evaluation. *The Teaching Professor* 4: 5–7.

Feldman, K.A. (1992). College students' views of male and female college teachers: Part I—evidence from the social laboratory and experiments. *Research in Higher Education* 33: 317–375.

Feldman, K.A. (1993). College students' views of male and female college teachers: Part II—evidence from students' evaluations of their classroom teachers. *Research in Higher Education* 34: 151–211.

Feldman, K.A. (1994). *Identifying Exemplary Teaching: Evidence from Course and Teacher Evaluations*. Paper commissioned by the National Center on Postsecondary Teaching, Learning, and Assessment for presentation at the Second AAHE Conference on Faculty Roles and Rewards.

Feldman, K.A. (forthcoming). Identifying exemplary teaching: Using data from course and teacher evaluations. In M.D. Svinicki and R.J. Menges (eds.), *Honoring Exemplary Teaching* (New Directions for Teaching and Learning). San Francisco: Jossey-Bass.

Feldman, K.A., and Newcomb, T.M. (1969). *The Impact of College on Students*. San Francisco: Jossey-Bass.

Feldman, K.A., and Paulsen, M.B. (eds.) (1994). *Teaching and Learning in the College Classroom*. Needham Heights, MA: Ginn Press.

French-Lazovik, G. (1974). Predictability of students' evaluation of college teachers from component ratings. *Journal of Educational Psychology* 66: 373–385.

Frey, P.W. (1973). Student ratings of teaching: Validity of several rating factors. *Science* 182: 83–85.

Frey, P.W. (1976). Validity of student instructional ratings as a function of their timing. *Journal of Higher Education* 47: 327–336.

Frey, P.W., Leonard, D.W., and Beatty, W.W. (1975). Student ratings of instruction: Validation research. *American Educational Research Journal* 12: 435–444.

Garber, H., 1964. *Certain Factors Underlying the Relationship between Course Grades and Student Judgments of College Teachers*. Unpublished doctoral dissertation, University of Connecticut, Storrs.

Gigliotti, R.J., and Buchtel, F.S. (1990). Attributional bias and course evaluation. *Journal of Educational Psychology* 82: 341–351.

Good, K.C. (1971). *Similarity of Student and Instructor Attitudes and Student's Attitudes Toward Instructors*. Unpublished doctoral dissertation, Purdue University, West Lafayette.

Greenwood, G.E., Hazelton, A., Smith, A.B., and Ware, W.B. (1976). A study of the validity of four types of student ratings of college teaching assessed on a criterion of student achievement gains. *Research in Higher Education* 5: 171–178.

Grush, J.E., and Costin, F. (1975). The student as consumer of the teaching process. *American Educational Research Journal* 12: 55–66.

Harry, J., and Goldner, N.S. (1972). The null relationship between teaching and research. *Sociology of Education* 45: 47–60.

Harvey, J.N., and Barker, D.G. (1970). Student evaluation of teaching effectiveness. *Improving College and University Teaching* 18: 275–278.

Hativa, N., and Raviv, A. (1993). Using a single score for summative teacher evaluation by students. *Research in Higher Education* 34: 625–646.

Hoffman, R.G. (1978). Variables affecting university student ratings of instructor behavior. *American Educational Research Journal* 15: 287–299.

Hoyt, D.P. (1973). Measurement of instructional effectiveness. *Research in Higher Education* 1: 367–378.

Jiobu, R.M., and Pollis, C.A. (1971). Student evaluations of courses and instructors. *American Sociologist* 6: 317–321.

King, P.M., and Kitchener, K.S. (1994). *Developing Reflective Judgment: Understanding and Promoting Intellectual Growth and Critical Thinking in Adolescents and Adults*. San Francisco: Jossey-Bass.

Kulik, M.A., and McKeachie, W.J. (1975). The evaluation of teachers in higher education. In F.N. Kerlinger (ed.), *Review of Research in Education* (Vol. 3). Itasca, II: F.E. Peacock.

Land, M.L. (1979). Low-inference variables of teacher clarity: Effects on student concept learning. *Journal of Educational Psychology* 71: 795–799.

Land, M.L. (1981). Actual and perceived teacher clarity: Relations to student achievement in science. *Journal of Research in Science Teaching* 18: 139–143.

Land, M.L., and Combs, A. (1981). *Teacher Clarity, Student Instructional Ratings, and Student Performance*. Paper read at the annual meeting of the American Educational Research Association.

Land, M.L., and Combs, N. (1982). Teacher behavior and student ratings. *Educational and Psychological Research* 2: 63–68.

Land, M.L., and Smith, L.R. (1979). The effect of low inference teacher clarity inhibitors and student achievement. *Journal of Teacher Education* 30: 55–57.

Land, M.L., and Smith, L.R. (1981). College student ratings and teacher behavior: An Experimental Study. *Journal of Social Studies Research* 5: 19–22.

Leftwich, W.H., and Remmers, H.H. (1992). A comparison of graphic and forced-choice ratings of teaching performance at the college and university level. *Purdue Universities Studies in Higher Education* 92: 3–31.

Leventhal, L. (1975). Teacher rating forms: Critique and reformulation of previous validation designs. *Canadian Psychological Review* 16: 269–276.

Levinson-Rose, J., and Menges, R.L. (1981). Improving college teaching: A critical review of research. *Review of Educational Research* 51: 403–434.

L'Hommedieu, R., Menges, R.J., and Brinko, K.T. (1988). *The Effects of Student Ratings Feedback to College Teachers: A Meta-analysis and Review of Research*. Unpublished manuscript, Northwestern University, Center for the Teaching Professions, Evanston.

L'Hommedieu, R., Menges, R.J., and Brinko, K.T. (1990). Methodological explanations for the modest effects of feedback. *Journal of Educational Psychology* 82: 232–241.

Maas, J.B., and Owen, T.R. (1973). *Cornell Inventory for Student Appraisal of Teaching and Courses: Manual of Instructions*. Ithaca, NY: Cornell University, Center for Improvement of Undergraduate Education.

Marsh, H.W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology* 76: 707–754.

Marsh, H.W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research* 11: 253–388.

Marsh, H.W. (1991a). Multidimensional students' evaluation of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology* 83: 285–296.

Marsh, H.W. (1991b). A multidimensional perspective on students' evaluations of teaching effectiveness: A reply to Abrami and d'Apollonia. *Journal of Educational Psychology* 83: 416–421.

Marsh, H.W. (in press). Weighting for the right criterion in the IDEA system: Global and specific ratings of teaching effectiveness and their relation to course objectives. *Journal of Educational Psychology*.

Marsh, H.W., and Dunkin, M.J. (1992). Students' evaluations of university teaching: A multidimensional approach. In J.C. Smart (ed.), *Higher Education: Handbook of Theory and Research* (Vol. 8). New York: Agathon Press.

Marsh, H.W., Fleiner, H., and Thomas, C.S. (1975). Validity and usefulness of student evaluations of instructional quality. *Journal of Educational Psychology* 67: 833–839.

Marsh, H.W., and Overall, J.U. (1980). Validity of students' evaluations of teaching effectiveness: Cognitive and affective criteria. *Journal of Educational Psychology* 72: 468–475.

McKeachie, W.J. (1979). Student ratings of faculty: A reprise. *Academe* 65: 384–397.
McKeachie, W.J. (1987). Instructional evaluation: Current issues and possible improvements. *Journal of Higher Education* 58: 344–350.
McKeachie, W.J., Lin, Y-G, and Mann, W. (1971). Student ratings of teacher effectiveness: Validity studies. *American Educational Research Association* 8: 435–445.
Miller, R.I. (1972). *Evaluating Faculty Performance*. San Francisco: Jossey-Bass.
Miller, R.I. (1974). *Developing Programs for Faculty Evaluation*. San Francisco: Jossey-Bass.
Mintzes, J.J., (1976–77). Field test and validation of a teaching evaluation instrument: The Student Opinion Survey of Teaching (A report submitted to the Senate Committee for Teaching and Learning, Faculty Senate, University of Windsor). Windsor, ON: University of Windsor.
Morgan, W.D., and Vasché, J.D. (1978). An Educational Production Function Approach to Teaching Effectiveness and Evaluation. *Journal of Economic Education* 9: 123–126.
Morsh, J.E., Burgess, G.G., and Smith, P.N. (1956). Student achievement as a measure of instructor effectiveness. *Journal of Educational Psychology* 47: 79–88.
Murray, H.G. (1980). *Evaluating University Teaching: A Review of Research*. Toronto: Ontario Confederation of University Faculty Associations.
Murray, H.G. (1983). Low-inference classroom teaching behaviors in relation to six measures of college teaching effectiveness. Proceedings of the Conference on the Evaluation and Improvement of University Teaching: The Canadian Experience (pp. 43–73). Montreal: McGill University, Centre for Teaching and Learning Service.
Murray, H.G. (1991). Effective teaching behaviors in the college classroom. In J.C. Smart (ed.), *Higher Education: Handbook of Theory and Research* (Vol. 7). New York: Agathon Press.
Orpen, C. (1980). Student evaluations of lecturers as an indicator of instructional quality: A validity study. *Journal of Educational Research* 74: 5–7.
Owen, P.H. (1967). *Some Dimensions of College Teaching: An Exploratory Study Using Critical Incidents and Factor Analyses of Student Ratings*. Unpublished doctoral dissertation, University of Houston, Houston.
Pascarella, E.T., and Terenzini, P.T. (1991). *How College Affects Students: Findings and Insights from Twenty Years of Research*. San Francisco: Jossey-Bass.
Perry, R.P. (1991). Perceived control in college students: Implications for instruction in higher education. In J.C. Smart (ed.), *Higher Education: Handbook of Theory and Research* (Vol. 7). New York: Agathon Press.
Plant, W.T., and Sawrey, J.M. (1970). Student ratings of psychology professors as teachers and the research involvement of the professors rated. *The Clinical Psychologist* 23: 15–16, 19.
Rankin, E.F., Jr., Greenmun, R., and Tracy, R.J. (1965). Factors related to student evaluations of a college reading course. *Journal of Reading* 9: 10–15.
Remmers, H.H. (1929). The college professor as the student sees him. *Purdue University Studies in Higher Education* 11: 1–63.
Remmers, H.H., Martin, F.D., and Elliott, D.N. (1949). Are students' ratings of instructors related to their grades? *Purdue University Studies in Higher Education* 66: 17–26.
Remmers, H.H., and Weisbrodt, J.A. (1964). *Manual of Instructions for Purdue Rating Scale of Instruction*. West Lafayette, IN: Purdue Research Foundation.

Rosenshine, B., Cohen, A., and Furst, N. (1973). Correlates of student preference ratings. *Journal of College Student Personnel* 14: 269–272.

Rubinstein, J., and Mitchell, H. (1970). Feeling free, student involvement, and appreciation. *Proceedings of the 78th Annual Convention of the American Psychological Association* 5: 623–624.

Sagen, H.B. (1974). Student, faculty, and department chairmen ratings of instructors: Who agrees with whom? *Research in Higher Education* 2: 265–272.

Sanders, J.A., and Wiseman, R.L. (1990). The effects of verbal and nonverbal teacher immediacy on perceived cognitive, affective, and behavioral learning in the multicultural classroom. *Communication Education* 39: 341–353.

Seldin, P. (1991). *The Teaching Portfolio*. Boston: Anker Publishing.

Shore, B.M., and associates (1986). *The Teaching Dossier: A Guide to Its Preparation and Use* (Rev. Ed.). Montreal: Canadian Association of University Teachers.

Smith, L.R., and Land, M.L. (1980). Student perception of teacher clarity in mathematics. *Journal for Research in Mathematics Education* 11: 137–146.

Sockloff, A.L. (1973). Instruments for student evaluation of faculty: Ideal and actual. In A.L. Sockloff (ed.), *Proceedings of the First Invitational Conference on Faculty Effectiveness as Evaluated by Students*. Philadelphia, PA: Temple University, Measurement and Research Center.

Solomon, D., Rosenberg, L., and Bezdek, W.E. (1964). Teacher behavior and student learning. *Journal of Educational Psychology* 55: 23–30.

Spencer, R.E. (1967). Analysis of the Instructor Rating Form—General Engineering Department. (Research Report No. 253). Urbana, II: University of Illinois, Measurement and Research Division, Office of Instructional Resources.

Stumpf, S.A., and Freedman, R.D. (1979). Expected grade covariation with student ratings of instruction: Individual versus class effects. *Journal of Educational Psychology* 71: 293–302.

Theall, M., Franklin, J., and Ludlow, L. (1990a). Attributions and retributions: Student ratings and the perceived causes of performance. *Instructional Evaluation* 11: 12–17.

Theall, M., Franklin, J., and Ludlow, L. (1990b). *Attributions or Retributions: Student Ratings and the Perceived causes of Performance*. Paper presented at the annual meeting of the American Educational Research Association.

Turner, R.L. (1970). Good teaching and its contexts. *Phi Delta Kappan* 52: 155–158.

Turner, R.L., and Thompson, R.P. (1974). *Relationships between College Student Ratings of Instructors and Residual Learning*. Paper presented at the annual meeting of the American Educational Research Association.

Van Horn, C. *An Analysis of the 1968 Course and Instructor Evaluation Report*. (Institutional Research Bulletin No. 2–68). West Lafayette, IN: Purdue University, Measurement and Research Center.

Walker, B.D. (1968). *An Investigation of Selected Variables Relative to the Manner in which a Population of Junior College Students Evaluate their Teachers*. Unpublished doctoral dissertation, University of Houston,

Widlak, F.W., McDaniel, E.D., and Feldhusen, J.F. (1973). *Factor Analysis of an Instructor Rating Scale*. Paper presented at the annual meeting of the American Educational Research Association.

Williams, H.Y., Jr. (1965). *College Students' Perceptions of the Personal Traits and Instructional Procedures of Good and Poor Teachers*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

APPENDIX

This appendix, with its listing of 28 instructional dimensions, first appeared in Feldman (1989b) in a slightly different version. For each of the instructional dimensions, examples of evaluation items that would be classified into it are given. For refinements and modifications to this list of dimensions and attendant coding scheme, see d'Apollonia, Abrami and Rosenfield (1993) and Abrami, d'Apollonia and Rosenfield (1996).

No. 1 *Teacher's Stimulation of Interest in the Course and Its Subject Matter*: "the instructor puts material across in an interesting way"; "the instructor gets students interested in the subject"; "it was easy to remain attentive"; "the teacher stimulated intellectual curiosity"; etc.

No. 2 *Teacher's Enthusiasm (for Subject or for Teaching)*: "the instructor shows interest and enthusiasm in the subject"; "the instructor seems to enjoy teaching"; "the teacher communicates a genuine desire to teach students"; "the instructor never showed boredom for teaching this class"; "the instructor shows energy and excitement"; etc.

No. 3 *Teacher's Knowledge of Subject Matter*: "the instructor has a good command of the subject material"; "the teacher has a thorough knowledge, basic and current, of the subject"; "the instructor has good knowledge about or beyond the textbook"; "the instructor knows the answers to questions students ask"; "the teacher keeps lecture material updated"; etc.

No. 4 *Teacher's Intellectual Expansiveness (and Intelligence)*: "the teacher is well informed in all related fields"; "the teacher has respect for other subject areas and indicates their relationship to his or her own subject of presentation"; "the teacher exhibited a high degree of cultural attainment"; etc.

No. 5 *Teacher's Preparation; Organization of the Course*: "the teacher was well prepared for each day's lecture"; "the presentation of the material is well organized"; the overall development of the course had good continuity"; "the instructor planned the activities of each class period in detail"; etc.

No. 6 *Clarity and Understandableness*: "the instructor made clear explanations"; the instructor interprets abstract ideas and theories clearly"; "the instructor makes good use of examples and illustrations to get across difficult points"; "the teacher effectively synthesizes and summarizes the material"; "the teacher answers students' questions in a way that helps students to understand"; etc.

No. 7 *Teacher's Elocutionary Skills*: "the instructor has a good vocal delivery"; "the teacher speaks distinctly, fluently and without hesitation"; "the teacher varied the speech and tone of his or her voice"; "the teacher has the ability to speak distinctly and be clearly heard"; "the instructor changed pitch, volume, or quality of speech"; etc.

No. 8 *Teacher's Sensitivity to, and Concern with, Class Level and Progress*: "the teacher was skilled in observing student reactions"; "the teacher was aware when students failed to keep up in class"; "the instructor teaches near the class level"; "the teacher takes an active personal interest in the progress of the class and shows a desire for students to learn"; etc.

No. 9 *Clarity of Course Objectives and Rdequirements*: "the purposes and policies of the course were made clear to the student"; "the instructor gave a clear idea of the student requirements"; "the teacher clearly defined student responsibilities in the course"; "the teacher tells students which topics are most important and what they can expect on tests"; "the instructor gave clear assignments"; etc.

No. 10 *Nature and Value of the Course Material (Including Its Usefulness and Relevance)*: "the teacher has the ability to apply material to real life"; "the instructor makes the course practical"; "there is worthwhile and informative material in lectures that doesn't duplicate the text"; "the course has excellent content"; "the class considers what we are learning worth learning"; etc.

No. 11 *Nature and Usefulness of Supplementary Materials and Teaching Aids*: "the homework assignments and supplementary readings were helpful in understanding the course"; "the teacher made good use of teaching aids such as films and other audio-visual materials"; "the instructor provided a variety of activities in class and used a variety of media (slides, films, projections, drawings) and outside resource persons"; etc.

No. 12 *Perceived Outcome or Impact of Instruction*: "gaining of new knowledge was facilitated by the instructor"; "I developed significant skills in the field"; "I developed increased sensitivity and evaluative judgment"; "the instructor has given me tools for attacking problems"; "the course has increased my general knowledge"; "apart from your personal feelings about the teacher, has he/she been instrumental in increasing knowledge of the course's subject matter"; etc.

No. 13 *Teacher's Fairness; Impartiality of Evaluation of Students; Quality of Examinations*: "grading in the course was fair"; "the instructor has definite standards and is impartial in grading"; "the exams reflect material emphasized in the course"; "test questions were clear"; "coverage of subject matter on exams was comprehensive"; etc.

No. 14 *Personality Characteristics ("Personality") of the Teacher*: "the teacher has a good sense of humor"; "the teacher was sincere and honest"; "the teacher is highly personable at all times in dress, voice, social grace, and manners"; "the instructor was free of personal peculiarities"; "the instructor is not autocratic and does not try to force us to accept his ideas and interpretations"; "the teacher exhibits a casual, informal attitude"; "the instructor laughed at his own mistakes"; etc.

No. 15 *Nature Quality, and Frequency of Feedback from the Teacher to Students*: "the teacher gave satisfactory feedback on graded material"; "criticism of papers was helpful to students"; "the teacher told students when they had done a good job"; "the teacher is prompt in returning tests and assignments"; etc.

No. 16 *Teacher's Encouragement of Questions and Discussion, and Openness to Opinions of Others*: "students felt free to ask questions or express opinions"; the instructor stimulated class discussions"; "the teacher encouraged students to express differences of opinions and to evaluate each other's ideas"; "the instructor invited criticisms of his or her own ideas"; "the teacher appeared receptive to new ideas and the viewpoints of others"; etc.

No. 17 *Intellectual Challenge and Encouragement of Independent Thought (by the Teacher and the Course)*: "this course challenged students intellectually"; "the teacher encouraged students to think out answers and follow up ideas"; "the teacher

attempts to stimulate creativity"; "the instructor raised challenging questions and problems"; etc.

No. 18 *Teacher's Concern and Respect for Students; Friendliness of the Teacher*: "the instructor seems to have a genuine interest in and concern for students"; "the teacher took students seriously"; "the instructor established good rapport with students"; "the teacher was friendly toward all students"; etc.

No. 19 *Teacher's Availability and Helpfulness*: "the instructor was willing to help students having difficulty"; "the instructor is willing to give individual attention"; "the teacher was available for consultation"; "the teacher was accessible to students outside of class"; etc.

No. 20 *Teacher Motivates Students to Do Their Best; High Standard of Performance Required*: "Instructor motivates students to do their best work"; "the instructor sets high standards of achievement for students"; "the teacher raises the aspirational level of students"; etc.

No. 21 *Teacher's Encouragement of Self-Initiated Learning*: "Students are encouraged to work independently"; "students assume much responsibility for their own learning"; "the general approach used in the course gives emphasis to learning on the students' own"; "the teacher does not suppress individual initiative"; etc.

No. 22 *Teacher's Productivity in Research Related Activities*: "The teacher talks about his own research"; "instructor displays high research accomplishments"; "the instructor publishes material related to his subject field"; etc.

No. 23 *Difficulty of the Course (and Workload)—Description*: "the workload and pace of the course was difficult"; "I spent a great many hours studying for this course"; "the amount of work required for this course was very heavy"; "this course required a lot of time"; "the instructor assigned very difficult reading"; etc.

No. 24 *Difficulty of the Course (and Workload)—Evaluation*: "the content of this course is too hard"; "the teacher's lectures and oral presentations are 'over my head' "; "the instructor often asked for more than students could get done"; "the instructor attempted to cover too much material and presented it too rapidly"; etc.

No. 25 *Classroom Management*: "the instructor controls class discussion to prevent rambling and confusion"; "the instructor maintained a classroom atmosphere conducive to learning"; "students are allowed to participate in deciding the course content"; "the teacher did not 'rule with an iron hand' "; etc.

No. 26 *Pleasantness of Classroom Atmosphere*: "the class does not make me nervous"; "I felt comfortable in this class"; "the instructor created an atmosphere in which students in the class seemed friendly"; "this was not one of those classes where students failed to laugh, joke, smile or show other signs of humor"; "the teacher is always criticizing and arguing with students"; etc.

No. 27 *Individualization of Teaching*: "instead of expecting every student to do the same thing, the instructor provides different activities for different students"; "my grade depends primarily upon my improvement over my past performance"; "in this class each student is accepted on his or her own merits"; "my grade is influenced by what is best for me as a person as well as by how much I have learned"; "the instructor evaluated each student as an individual"; etc.

No. 28 *Teacher Pursued and/or Met Course Objectives*: "the instructor accomplished what he or she set out to do"; "there was close agreement between the announced objectives of the course and what was actually taught"; "course objectives stated agreed with those actually pursued"; etc.

COMMENTARY AND UPDATE ON FELDMAN'S (1997) "IDENTIFYING EXEMPLARY TEACHERS AND TEACHING: EVIDENCE FROM STUDENT RATINGS"

Michael Theall* **and Kenneth A. Feldman†**

**Youngstown State University*
mtheall@ysu.edu
†SUNY-Stony Brook

Abstract

In the original chapter (1997), Feldman explores how student ratings can be used to identify those teachers who are seen by students as exemplary, while noting certain precautions (which involve myths, half-truths and bias) in doing so. He also analyzes how exemplary teaching itself can be identified in terms of specific pedagogical dispositions, behaviors and practices of teachers. While the essential findings of this earlier analysis remain valid, there have been changes in the nature and focus of research on college teaching and its evaluation. As well, new challenges and developments are forcing higher education to rethink its paradigms and practices in such areas as teaching, the evaluation of faculty performance, and the kinds of support faculty need to meet the increasingly complex professional demands placed on the professoriate. The co-authors of the commentary and update (Theall and Feldman) review the principal findings of the original chapter, discuss the literature of the past decade, and offer suggestions for ways in which higher education and the professoriate can survive and flourish in the future

Key Words: College teaching; dimensions of teaching; exemplary teaching; student ratings of instruction; reliability and validity; myths vs. research evidence; faculty evaluation; the professoriate; higher education; paradigm shift; research and practice; faculty development; professional enrichment; faculty as meta-professionals; faculty careers

Reviewing the extant literature, Feldman (1997) explored how student ratings could be used to identify those persons who are seen by students as exemplary teachers, while noting certain precautions (involving current myths and half-truths as well as issues of bias) in doing so. He then analyzed how exemplary teaching itself can be identified in terms of specific pedagogical dispositions, behaviors and practices of teachers. He reviewed dimensions of teaching that are associated with student learning and with overall evaluations of teachers.

Since Feldman's chapter appeared, the number of publications about student ratings has not noticeably diminished nor has the amount of discussion abated. Although, in general, the major conclusions of his chapter still hold, two tracks of activity have become apparent—both of which we consider in various places in this commentary and update. One track is the continuation of scholarship by researchers and practitioners in the field with an increasing emphasis on bringing the research into practice. This activity has not gone unnoticed as evidenced by the fact that the 2005 American Educational Research Association's "Interpretive Scholarship, Relating Research to Practice Award" went to evaluation and ratings researchers doing just this kind of work. The second track has been less productive in terms of improving practice. It is represented by an array of opinion and reports of investigations attempting to prove that ratings are biased, are the cause of grade inflation, and are threats to promotion, tenure, academic freedom, and the general quality of higher education. One result of this activity has been the extension of misinformation and mythology surrounding ratings, which in effect has made improved practice more difficult (Aleamoni, 1987; Feldman, 1997).

GENERALIZATIONS AND CONCLUSIONS ABOUT RATINGS AS EVIDENCE OF EXEMPLARY TEACHERS AND TEACHING: SOME CAUTIONS

Feldman cautioned that his 1997 chapter was "...not intended as a comprehensive review of the research literature on evaluation of college students (ratings) of their teachers or on the correlates of effective teaching in college." (p. 385, parenthetical term added). This caution still applies for four reasons. First, it is clear that the number and variety of issues affecting teaching and learning is exceptionally large and complex, and thus beyond the scope of the present update. For example, recent work by Pascarella and Terenzini (2005) and Kuh et al. (2005) demonstrates that student performance is affected by a number of conditions beyond classroom teaching and other efforts of the faculty members. College instructors, existing in this same set of conditions, cannot help but be influenced as well, and thus their satisfaction as well as that of their students can affect their teaching and their students' perceptions of it (Cranton & Knoop, 1991). Ratings reflect students' opinions about teaching, and they do correlate with learning (Cohen, 1981), but to some degree they also indicate students' general satisfaction with their experiences. Environmental factors can affect those

experiences and thus complicate an already complex measurement situation. Indeed, the complexity of the whole teaching-learning picture demands a broader view that both includes and goes beyond ratings as evidence of exemplary teaching.

A second reason for repeating the caveat is that while Feldman's chapter can remain essentially unchallenged in terms of its conclusions (because there is little in the way of substantial new evidence that contradicts his interpretations), at the same time there is an absence of new literature about exemplary teaching in contexts essentially nonexistent when the earlier analysis was completed. Of particular note, for example, is the growth of technologically enhanced instruction and on-line or other "distance" teaching and learning experiences. Thus, Feldman's (1989) work refining and extending the synthesis of data from multivalidation studies (reviewed in Feldman's 1997 chapter) remains a primary source of information about the dimensions of college teaching in traditional teaching and learning settings (also, see Abrami, d'Apollonia and Rosenfield, 1996). But, the 1997 chapter cannot be updated without also considering Feldman's (1998) chapter urging readers to consider the effects of context and "unresolved issues."

The growth of instruction in settings other than traditional classrooms raises questions about the extent to which established models and psychometric techniques can be transplanted into these new situations. Because there has not been a great deal of research on how to evaluate teaching in these contexts, these questions remain unresolved. Using the same traditional questionnaires and producing the same traditional reports raises serious validity issues. In addition, and given the number of references found in opinion pieces in the press and elsewhere, the emergence of private or for-profit on-line ratings has added to the faculty's legitimate concern about the misuse of ratings data. Clearly, the issues related to technological innovations are numerous, and while we note their impact here we cannot explore them in depth.

The third reason involves the nature and variety of publications specifically on student ratings. Earlier literature contained many in-depth analyses of ratings issues characterized by reports drawn from validation studies (e.g., Cashin, 1990, with IDEA; Centra, 1972, with SIR; Marsh, 1987, with SEEQ), syntheses or meta-analyses (e.g., Cohen, 1981; Feldman, 1989; Abrami, d'Apollonia and Rosenfield, 1996), the use of large databases to explore specific issues (e.g., Franklin and Theall, 1992, with TCEP), the involvement of scientists/researchers whose primary research emphases were teaching, learning, or ratings

themselves (e.g., Doyle, 1975; Centra, 1979; Marsh, 1984) and, importantly, the extended discussion of reported results. An example of this last factor can be found in commentary and studies following the Naftulin, Ware and Donnelly (1973) "Dr. Fox" article. Commentaries were published in four issues of *Instructional Evaluation* between 1979 and 1982,[1] and Raymond Perry and associates conducted a series of studies on educational seduction and instructor expressiveness between 1979 and 1986, incorporating perceived control as a major variable in their later work.[2] Though there was public notice of the "Dr. Fox" study, the primary participants in the dialogue were the researchers themselves. This is less the case today, as almost any opinion from any quarter (it would seem) is deemed worthy of publication, and because communications technologies allow anyone to publish and widely circulate an opinion without the process required in traditional refereed environments.

Finally, affecting the scope of this update is the general descent of the status of the professoriate and higher education itself. Even respected academicians have produced work with clear sarcasm in their titles—for example, "Dry Rot in the Ivory Tower" (Campbell, 2000) and "Declining by Degrees" (Hersh and Merrow, 2005). A spate of books and opinions has inflamed the general public, editorial writers, and legislators who feel ever more comfortable demanding "accountability." The interesting irony is that if, to the joy of many critics, ratings were to be eliminated in favor of student learning as a measure of teaching excellence, then the same arguments used to question the reliability and validity of ratings would arise with respect to testing and grading. "Grade inflation," which according to these same critics (e.g., Johnson, 2003; Trout, 2000) is the by-product of ratings, would not disappear. Rather, grades might either become favored by faculty as evidence of teaching excellence ("Look how well I did. My students all got As!") or they would become the new criteria by which poor teaching would be characterized ("S/he must be a poor teacher! Look how many students got As). Thus, in this brave new world, teachers might

[1] *Instructional Evaluation* (now *Instructional Evaluation and Faculty Development*) is a semi-annual publication of the Special Interest Group in Faculty Teaching, Evaluation, and Development of the American Educational Research Association. Issues from 1996 are available on-line at: http://www.umanitoba.ca/uts/sigfted/backissues.php. Earlier issues can be purchased using information provided at: http://www.umanitoba.ca/uts/sigfted/iefdi/spring00/bkissues.htm.

[2] Studies with instructor expressiveness as a variable include Perry, Abrami and Leventhal (1979) through Perry, Magnusson, Parsonson and Dickens (1986). The conclusions of the research were that expressiveness alone does not enhance achievement but can influence ratings of specific presentation skills, and that in combination with appropriate content it can positively influence both ratings and achievement.

provide evidence of excellence by either "dumbing down" courses to maximize the numbers of As or, in the opposite perversion, failing as many students as possible.

THE PUBLIC DEBATE ON STUDENT RATINGS

Discussion of ratings issues has continued, and perhaps has even been unduly influenced by recent publications. For example, Johnson (2003) has supported a proposal at Duke University (see Gose, 1997) whereby class grade profiles would be used to rate classes so that the grades students received could then be given more or less weight in a calculation of the GPA. Such reaction and over-reaction seems based primarily on assumptions that grade inflation has been caused by the use of ratings and by a focus on learners as customers or consumers. The language of complaints almost always includes these issues in a simplistic way without crediting established findings (e.g., Cohen, 1981; Marsh, 1987) or taking account of the larger picture of improving evaluation and teaching in complimentary ways (e.g., Theall & Franklin, 1991).

Many recent publications are based on one-time and/or small-sample studies that vary substantially from methodologically accepted practice (e.g., Williams & Ceci, 1997), many include ratings issues in work from other disciplinary perspectives (e.g., Hamermesh & Parker, 2003), many are more opinion pieces than specific research on ratings (e.g., Trout, 2000), and few are by researchers whose primary emphasis has been faculty evaluation or student ratings (which would include all of the above-cited items). Many of these pieces have become well known by virtue of the interest of widely distributed publications (e.g., *Academe*, *The Chronicle of Higher Education*) in the controversy surrounding ratings.

One partial exception was the substantial work by Greenwald and Gillmore (1997a, 1997b) that culminated in a "Current Issues" section of *American Psychologist* (Vol. 52, No. 11) devoted to the topic of grade inflation and ratings. That was followed by an AERA symposium on the same topic. While there was considerable disagreement with Greenwald and Gillmore's contention that grading leniency was a "contaminant" of ratings to be statistically corrected, the point is that the work included a series of studies using a substantial database, and it was followed by an extended debate on the work and its conclusions by experienced ratings researchers. Nonetheless, the *Chronicle* published a lengthy article (Wilson, 1998) that contained errors serious enough to attract

critical letters from many researchers who were quoted, including Gerald Gillmore himself.

This over-emphasis on criticisms of ratings has led to another problem: the propagation of the criticisms themselves as a separate and widely believed mythology about ratings. Arreola (2005a) maintains that these myths have become a kind of "...common knowledge, so pervasive that it far overshadows the 'truth' concerning student ratings and other faculty evaluation tools buried in the pages of psychometric journals" (p. 1). This pattern has even progressed to the point where writers critical of ratings (e.g., Johnson, 2003) refer to well-established and often-replicated ratings findings as "myths." Not surprisingly, there has been criticism of Johnson's book from within the community of ratings researchers and practitioners (e.g., Perry, 2004). One implication of Johnson's notoriety is that experienced evaluation and ratings researchers need to do a better job of putting their findings before two critical audiences: the faculty and administrators who use these data (Arreola, 2005a, 2005b).

A POSITIVE SHIFT IN EMPHASIS: RESEARCH INFORMING PRACTICE

Apart from the public debate on student ratings, there has been a shift in emphasis in recent years in the study and consideration of student ratings. Ratings researchers, writers, and practitioners have tended to move from necessary but sometimes narrow psychometric investigations concerned with validity and reliability of ratings, to the application of evaluation and ratings research to practice. Beginning as early as Theall and Franklin (1990a), through updates of previous work, this pattern has led to detailed descriptions of, and guidelines for, the development of "comprehensive evaluation systems" (Arreola, 2000). As recently as the 2005 meeting of the American Educational Research Association, the question, "Valid Faculty Evaluation Data: Are There Any?" was raised with an emphasis on improving evaluation practice rather than establishing or re-establishing the purely technical validity and reliability of ratings.

To a large extent this stream of thinking echoes Feldman's (1998) emphasis on the importance of a "continuing quest" (when analyzing the correlates and determinants of effective instruction) for "...establishing the conditions or contexts under which relationships become stronger or weaker...or change in some other way.... The quest calls attention to determining the importance of 'interaction effects' as well as 'main effects'" (p. 36). The context in which evaluation

takes place has been shown to have a potentially serious effect on the way that ratings data can be both interpreted and used. For example, Franklin and Theall (1993) found gender differences in ratings in certain academic departments. Although women had lower average ratings, further exploration showed that in those departments women had been disproportionately assigned to teach large, introductory, required classes—those where teachers in general might be expected to have lower ratings. Replication of the study at another institution where course assignments were equally distributed found no gender differences. The information available to faculty and administrators rarely includes analysis that goes beyond mean scores or averages; thus, contextual subtleties are lost, misinterpretation is more likely, and integration of ratings research with practice is hindered.

Another contextual factor that can influence ratings is institutional type as exemplified, say, by the different emphases and operations of community colleges and research universities described by Birnbaum (1988). Such differences can affect the perceptions of faculty, students, and administrators at these institutions, thus influencing the expectations held for faculty work and the definitions of "exemplary teaching." Contextual differences can further occur across disciplines in average ratings of teachers and courses (Cashin, 1990); in instructional choices of faculty (Franklin & Theall, 1992); in their effects on students' assimilation into the disciplines (Smart, Feldman, and Ethington, 2000); and in the extent to which teachers communicate expectations about course work (Franklin & Theall, 1995).

IMPROVING THE PRACTICE OF RATING TEACHERS AND INSTRUCTION

In the past half-dozen years or so, there have been several new attempts to improve ratings and evaluation practice. Perhaps the most focused work is by Arreola (2000), who describes a detailed process for "Developing a Comprehensive Faculty Evaluation System." Arreola outlines an eight-step process that can be used to generate institutional dialogue on issues that need to be discussed before any evaluation or ratings process is begun. Theall and Franklin (1990b) proposed that ratings are only one part of "complex evaluation systems," and Arreola's (2000) process outline remains the only articulated approach that takes into account and deals with the contextual issues that greatly influence evaluation and ratings practice on a campus-by-campus basis.

Arreola has not been alone in pressing for improved practice. Indeed, no less than six volumes of the Jossey Bass "New Directions" series have been devoted to ratings issues since Feldman's (1997) chapter was published. The first contribution to this extended discussion was from Ryan (2000), who proposed a "Vision for the Future" based not only on sound measurement, but on "…philosophical issues that need to be addressed if faculty evaluation is to receive the respect and attention it deserves" (backpage, "From the Editor"). Theall, Abrami and Mets (2001) asked about ratings, "Are they valid? How can we best use them?" Included in their volume are chapters reminiscent of the depth and extent of earlier exemplary dialogue and debate in the field (noted earlier in the present commentary). Lewis (2001) edited a volume concentrating on "Techniques and Strategies for Interpreting Student Evaluations." In particular, this set of articles connects issues of accountability to the faculty evaluation process and considers ratings as existing within the context of department, college, and institutional imperatives (and as needing to be responsive to these pressures). This volume was immediately followed by one (Knapper & Cranton, 2001) presenting "Fresh Approaches to the Evaluation of Teaching." Colbeck (2002) took an appropriately broad view of "Evaluating Faculty Performance," noting that "Forces for change within and outside academe are modifying faculty work and the way that work is—or should be—evaluated" (p. 1). Finally, Johnson and Sorenson (2004) presented a specific discussion of a new aspect of the ratings and evaluation picture: the rapidly increasing use of on-line systems. Acknowledging that this rapid growth is occurring "…even amidst challenges and doubt" (p.1), they and other contributors present a balanced review of the advantages and disadvantages of on-line systems.[3]

BEYOND RATINGS AS EVIDENCE OF EXEMPLARY TEACHING: ENHANCING FACULTY CAREERS

It can be argued that college teaching and learning, evaluations of teachers, and higher education itself have changed to the point where it is no longer reasonable or prudent to consider student ratings of

[3] Although there is space only to list references here, in the past yen years or so various volumes of *Higher Education: Handbook of Theory and Research* have published articles dealing with evaluation of teaching, teaching effectiveness, and improvement in instructional practices: see, for example, Boice (1997), Feldman (1998), Murray (2001), Cashin (2003), and Centra (2004).

teaching effectiveness without also considering the context in which they occur. These ratings are or should be embedded in processes (faculty evaluation and development) that are connected to department and college issues (staffing, funding, competition for resources), institutional issues (assessment, accreditation, reputation) and other matters that extend beyond campus (public support, legislation, accountability). A systematic and ecological approach to evaluation and ratings is needed because effective practice is not possible without consideration of (and coordination with) these other issues.

Recent work (Arreola, 2005b, 2005c; Arreola, Theall, & Aleamoni, 2003; Theall, 2002)[4] has expanded on past approaches and incorporated a wider view that encompasses past literature on faculty evaluation and development, the professoriate, business, science, communications, and reaction to change, as well as new discussions of how contemporary changes and forces are affecting higher education and the professoriate (e.g., Hersh & Merrow, 2005; Newman, Couturier, & Scurry, 2004).

Defining the professoriate as requiring in-depth expertise in a disciplinary or "base profession" as well as professional skills in several other "meta-professional" areas, Arreola, Theall and Aleamoni (2003) have developed a two-dimensional matrix that arrays four faculty roles (Teaching, Scholarly and Creative Activities, Service, and Administration) against three base-profession skills (content expertise, practice/clinical skills, and research techniques), and twenty meta-professional skill sets (e.g., instructional design skills, group process and team-building skills, public speaking skills). The frequency of need for each skill-by-role cell in the matrix is indicated by color-coding. Five additional matrices are provided, in which the roles are broken down into component parts. For example, the "teaching" role has seven contextual components ranging from traditional classroom situations to on-line and distance learning. Scholarly and Creative Activities, Service, and Administration are also broken down into their contextual components, and a special matrix demonstrates how Boyer's (1990) "scholarship of teaching (and learning)" presents a special case of meta-professional requirements.

[4] The "Meta-Profession Project" is an ongoing effort to improve practice in faculty evaluation and development. It is based on the analysis of the roles faculty are required to fill and the skills necessary to successfully carry out role responsibilities. The basic roles and skill sets are displayed in a series of two-dimensional matrices available at the project website at http://www.cedanet.com/meta. The site also contains an overview of the concept and project, copies of various papers and publications, and related information.

The conceptualization of the meta-profession and the matrices provide a framework for exploring the nature and demands of faculty work on a campus-by-campus basis and thus for improving practice in faculty evaluation and development. Similarly, this exploration can lead to improved policy and provide numerous opportunities to investigate faculty work on a broad scale, particularly as it is affected by variables such as institutional type, individual and campus demographics, and changes in prevailing economic and other conditions.

A FINAL COMMENT

Clearly, various psychometric issues (including reliability and validity) are important to the study and improvement of student ratings (Feldman, 1998). Even so, and despite these technical requirements, faculty evaluation and the use of student ratings involve more than psychometric issues; important professional, political, social, personnel, and personal issues also come into play. Recent years have seen the potential threat of a seemingly endless and unproductive debate on reliability and validity issues—unproductive in the sense that what has been established in over fifty years of substantial research has been largely ignored for reasons that include administrative convenience, ignorance, personal biases, suspicion, fear, and the occasional hostility that surrounds any evaluative process.

Evaluation and student ratings are unlikely to improve until practice is based on a careful and accurate analysis of the work required of faculty, the skills required to do that work, and the levels of performance expected. Further, good practice requires the creation of complete systems that acknowledge the absolute need to blend professional and faculty development resources with those necessary for fair and equitable faculty evaluation. Student ratings form a part of this picture, but too often have been inappropriately employed with the result that there has been a disproportionate amount of attention, debate and dissension, accompanied by misinformation based on questionable research, a ratings mythology, and the propagation of a general sense that the use of ratings somehow demeans the teaching profession. To the extent that the negative feeling is based on a degree of truth that has its base in poor practice, then improving practice becomes a critical agenda.

References

Abrami, P.C., d'Apollonia, S., and Rosenfield, S. (1996). The dimensionality of student ratings of instruction: What we know and what we do not. In J.C. Smart (ed.), *Higher Education: Handbook of Theory and Research*. New York: Agathon Press.

Aleamoni, L.M. (1987). Student rating myths versus research facts. *Journal of Personnel Evaluation in Education* 1: 111–119.

Arreola, R.A. (2000). *Developing a Comprehensive Faculty Evaluation System* (2nd edition). Bolton, MA: Anker Publishing Company.

Arreola, R.A. (2005a). Validity, like beauty is…Paper presented at the annual meeting of the American Educational Research Association. (Available at: http://www.cedanet.com/meta/ and http://www.umanitoba.ca/uts/sigfted/backissues.php)

Arreola, R.A. (2005b). Crossing over to the dark side: Translating research in faculty evaluation into academic policy and practice. Invited address presented at the annual meeting of the American Educational Research Association. (Available at: http://www.cedanet.com/meta/)

Arreola, R.A. (2005c). The monster at the foot of the bed. *To Improve the Academy* 24: 15–28.

Arreola, R.A., Theall, M., and Aleamoni, L.M. (2003). Beyond scholarship: Recognizing the multiple roles of the professoriate. Paper presented at the annual meeting of the American Educational Research Association. (Available at: http://www.cedanet.com/meta/)

Birnbaum, R. (1988). *How Colleges Work*. San Francisco: Jossey Bass.

Boice, B. (1997). What discourages research-practioners in faculty development. In J.C. Smart (ed.), *Higher Education: Handbook of Theory and Research* (Vol. 12). New York: Agathon Press.

Boyer, E.L. (1990). *Scholarship Reconsidered*. San Francisco: Jossey Bass.

Campbell, J.R. (2000). *Dry Rot in the Ivory Tower*. Lanhan, MD: University Press of America.

Cashin, W.E. (1990). Students do rate different academic fields differently. 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.

Cashin, W.E. (2003). Evaluating college and university teaching: Reflections of a practioner. In J.C. Smart (ed.), *Higher Education: Handbook of Theory and Research* (Vol. 18). Norwell, MA: Kluwer Academic Publishers.

Centra, J.A. (1972). The student instructional report: Its development and uses. (Student Instructional Report No. 1). Princeton, NJ: Educational Testing Service.

Centra, J.A. (1979). *Determining Faculty Effectiveness: Assessing Teaching, Research, and Service for Personnel Decisions and Improvement*. San Francisco: Jossey Bass.

Centra, J.A. (2004). Service through research: My life in higher education. In J.C. Smart (ed.), *Higher Education: Handbook of Theory and Research* (Vol. 19). Norwell, MA: Kluwer Academic Publishers.

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.

Colbeck, C.L. (ed.) (2002). *Evaluating Faculty Performance* (New Directions for Institutional Research No. 114). San Francisco: Jossey Bass.

Cranton, P.A., and Knoop, R. (1991) Incorporating job satisfaction into a model of instructional effectiveness. In M. Theall and J. Franklin (eds.), *Effective Practices for Improving Teaching* (New Directions for Teaching and Learning No. 48). San Francisco: Jossey Bass.

Doyle, K.O. (1975). *Student Evaluation of Instruction*. Lexington, MA: D. C. Heath.

Feldman, K.A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education* 30: 583–645.

Feldman, K.A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R.P. Perry and J.C. Smart (eds.), *Effective Teaching in Higher Education: Research and Practice*. New York: Agathon Press.

Feldman, K.A. (1998). Reflections on the effective study of college teaching and student ratings: One continuing quest and two unresolved issues. In J.C. Smart (ed.), *Higher Education: Handbook of Theory and Research* (Vol. 13). New York: Agathon Press.

Franklin, J., and Theall, M. (1992). *Disciplinary differences, instructional goals and activities, measures of student performance, and student ratings of instruction*. Paper presented at the annual meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 346 786)

Franklin, J., and Theall, M. (1993). Student ratings of instruction and gender differences revisited. Paper presented at the annual meeting of the American Educational Research Association.

Franklin, J., and Theall, M. (1995). The relationship of disciplinary differences and the value of class preparation time to student ratings of instruction. In N. Hativa and M. Marincovich (eds.), *Disciplinary Differences in Teaching and Learning: Implication for Practice* (New Directions for Teaching and Learning No. 64). San Francisco: Jossey-Bass.

Gose, B. (1997). Duke may shift grading system to reward students who take challenging classes. *The Chronicle of Higher Education* February 14: A43.

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

Greenwald, A.G., and Gillmore, G.M. (1997b). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology* 89: 743–751.

Hamermesh, D., and Parker, A.M. (2003). Beauty in the classroom: Professors' pulchritude and putative pedagogical productivity. (NBER Working Paper No. W9853). Austin TX: University of Texas at Austin. Abstract available at: http://ssrn.com/abstract=425589

Hersh, R.H., and Merrow, J. (2005). *Declining by Degrees: Higher Education at Risk*. New York: Palgrave Macmillan.

Johnson, T., and Sorenson, L. (2004). *Online Student Ratings of Instruction* (New Directions for Teaching and Learning No. 96). San Francisco: Jossey Bass.

Johnson, V. (2003). *Grade Inflation: A Crisis in Higher Education*. New York: Springer Verlag.

Knapper, C., and Cranton, P.A. (eds.) (2001). *Fresh Approaches to the Evaluation of Teaching* (New Directions for Teaching and Learning No. 88). San Francisco: Jossey Bass.

Kuh, G.D., Kinzie, J., Schuh, J.H., Whitt, E.J., and Associates (2005). *Student Success in College: Creating Conditions That Matter*. 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.

Marsh, H.W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology* 76: 707–754.

Marsh, H.W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research* 11: 253–388.

Murray, H.G. (2001). Low-interence teaching behaviors and college teaching effectiveness: Recent developments and controversies. In J.C. Smart (ed.), *Higher Education: Handbook of Theory and Research* (Vol. 16). New York: Agathon Press.

Naftulin, D.H., Ware, J.E., and Donnelly, F.A. (1973). The Dr. Fox lecture: A paradigm of educational seduction. *Journal of Medical Education* 48: 630–635.

Newman, F., Couturier, L., and Scurry, J. (2004). *The Future of Higher Education: Rhetoric, Reality, and the Risks of the Market*. San Francisco: Jossey Bass.

Pascarella, E.T., and Terenzini, P.T. (2005). *How College Affects Students, Volume 2: A Third Decade of Research*. San Francisco: Jossey Bass.

Perry, R.P. (2004). Review of a V. Johnson's *Grade Inflation: A Crisis in College Education*. *Academe* January-February: 10–13.

Perry, R.P., Abrami, P.C., and Leventhal, L. (1979). Educational seduction: The effect of instructor expressiveness and lecture contrent on student ratings and achievement. *Journal of Educational Psychology* 71: 107–116.

Perry, R.P., Magnusson, J.L., Parsonson, K.L., and Dickens, W.J. (1986). Perceived control in the college classroom: Limitations in instructor expressiveness due to non-contingent feedback and lecture content. *Journal of Educational Psychology* 78: 96–107.

Ryan, K.E. (ed.) (2000). *Evaluating Teaching in Higher Education: A Vision of the Future* (New Directions for Teaching and Learning No. 83). San Francisco: Jossey Bass.

Smart, J.C., Feldman, K.A., and Ethington, C.A. (2000). *Academic Disciplines: Holland's Theory and the Study of College Students and Faculty*. Nashville, TN: Vanderbilt University Press.

Theall, M. (2002). Leadership in faculty evaluation and development: Some thoughts on why and how the *meta*-profession can control its own destiny. Invited address at the annual meeting of the American Educational Research Association. (Available at: http://www.cedanet.com/meta/)

Theall, M., Abrami, P.C., and 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., and Franklin, J.L. (eds.) (1990a). *Student Ratings of Instruction: Issues for Improving Practice* (New Directions for Teaching and Learning No. 43). San Francisco: Jossey Bass.

Theall, M., and Franklin, J.L. (1990b). 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.

Theall, M., and Franklin, J. (eds.) (1991). *Effective Practices for Improving Teaching* (New Directions for Teaching and Learning No. 48). San Francisco: Jossey Bass.

Trout, P. (2000). Flunking the test: The dismal record of student evaluations. *Academe* July-August: 58–61.

Williams, W.M., and Ceci, S.J. (1997). How'm I doing?: Problems with student ratings of instructors and courses. *Change* 29(5): 13–23.

Wilson, R. (1998). New research casts doubt on value of student evaluations of professors. *The Chronicle of Higher Education* January 16: A1, A16.