A review of the data on the effectiveness of student evaluations reveals that they are a widely used but greatly misunderstood source of information for exemplary teaching.

Identifying Exemplary Teaching: Using Data from Course and Teacher Evaluations

Kenneth A. Feldman

Many colleges and universities are currently putting increased emphasis on good teaching and on designating, honoring, and rewarding good teachers. Consequently, student ratings, long a staple for student course selection, instructor feedback, and administrative personnel decisions, are likely to be applied to this additional purpose. Yet, for all their use, student ratings of instructors and instruction are hardly universally accepted. Some college teachers believe ratings are not reliable, valid, or useful and may even be harmful. Others believe more or less the opposite.

If the credibility of teacher and course evaluations is to be based on more than mere opinion, we must consult the research on their use. In one of the best overviews of this research, after one hundred pages or so of careful, critical, and reflective analysis, Marsh (1987) concludes the following:

Research described in this article demonstrates that student ratings are clearly multidimensional, quite reliable, reasonably valid, relatively uncontaminated by

Note: This chapter is a much shortened version of a paper commissioned by the National Center on Postsecondary Teaching, Learning, and Assessment for presentation at the second AAHE Conference on Faculty Roles and Rewards, held in 1994. I would like to thank Robert Menges and Maryellen Weimer for their thoughtful suggestions for the original paper and Marilla Svinicki for her invaluable help in condensing it for the present volume. An expanded version, offering fuller analysis and greater detail, can be found under the title, "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, 1996, in press).
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 decisions, 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 chapter analyzes the differential importance of the individual items that constitute the rating forms used to evaluate teachers and courses and explores various interpretations that can be made of information gathered from students about their teachers.

Identifying Instructional Dimensions Important to Effective Teaching

Evaluation instruments try to capture the multidimensionality of teaching. These instruments, as Marsh and Dunkin (1992) point out, are typically constructed from "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 empirical approach that emphasizes statistical techniques such as factor analysis or multitrait-multimethod analysis. In my own dimensional analysis, I have recently extended an early set of roughly twenty instructional dimensions of teaching components to include twenty-eight dimensions (see Feldman, 1988, 1989a, 1993).

Relation Between Teaching and Student Learning. Based in part on work by d'Apollonia and Abrami (1987, 1988) and Abrami, Cohen, and d'Apollonia (1988), I extended Cohen's (1980b, 1981, 1987) meta-analysis on the association between student achievement and various instructional dimensions (Feldman, 1989a, 1990). The correlations (from this reanalysis) between specific survey items and student achievement are shown in Table 6.1 and range from +.57 to -.11.

Item ranks are shown in parentheses. All but one (for dimension 22) are positive, and all but three (for 22, 23, and 24) are statistically significant. The highest correlations of +.57 and +.56 are for dimensions 1 ("teacher's preparation and course organization") and 2 ("teacher's clarity and understandableness"). The "teacher's pursuit and/or meeting of course objectives" (dimension
Table 6.1. Importance of Instructional Dimensions Using Two Different Indicators of Importance

<table>
<thead>
<tr>
<th>Instructional Dimension</th>
<th>(1) Correlation with Student Achievement (Larger = More Important; Rank in Parentheses)</th>
<th>(2) Average Standardized Rank Based on Correlation with Overall Evaluation (Smaller = More Important; Rank in Parentheses)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Teacher's preparation; organization of course</td>
<td>.57 (1)</td>
<td>.41 (6)</td>
</tr>
<tr>
<td>2. Clarity and understandableness</td>
<td>.56 (2)</td>
<td>.25 (2)</td>
</tr>
<tr>
<td>3. Teacher pursued and/or met course objectives</td>
<td>.49 (NA)</td>
<td>NA</td>
</tr>
<tr>
<td>4. Perceived outcome or impact of instruction</td>
<td>.46 (3)</td>
<td>.28 (3)</td>
</tr>
<tr>
<td>5. Teacher's stimulation of interest in the course and its subject matter</td>
<td>.38 (4)</td>
<td>.20 (1)</td>
</tr>
<tr>
<td>6. Teacher motivates students to do their best; high standard of performance required</td>
<td>.38 (NA)</td>
<td>NA</td>
</tr>
<tr>
<td>7. Teacher's encouragement of questions and discussion, and openness to opinions of others</td>
<td>.36 (5.5)</td>
<td>.60 (11)</td>
</tr>
<tr>
<td>8. Teacher's availability and helpfulness</td>
<td>.36 (5.5)</td>
<td>.74 (16)</td>
</tr>
<tr>
<td>9. Teacher's elocutionary skills</td>
<td>.35 (7.5)</td>
<td>.49 (10)</td>
</tr>
<tr>
<td>10. Clarity of course objectives and requirements</td>
<td>.35 (7.5)</td>
<td>.45 (7)</td>
</tr>
<tr>
<td>11. Teacher's knowledge of the subject</td>
<td>.34 (9)</td>
<td>.48 (9)</td>
</tr>
<tr>
<td>12. Teacher's sensitivity to, and concern with, class level and progress</td>
<td>.30 (10)</td>
<td>.40 (5)</td>
</tr>
<tr>
<td>13. Teacher's enthusiasm (for subject or for teaching)</td>
<td>.27 (11)</td>
<td>.46 (8)</td>
</tr>
<tr>
<td>14. Teacher's fairness; impartiality of evaluation of students; quality of examinations</td>
<td>.26 (12)</td>
<td>.72 (14.5)</td>
</tr>
<tr>
<td>15. Teacher's classroom management</td>
<td>.26 (NA)</td>
<td>NA</td>
</tr>
<tr>
<td>16. Intellectual challenge and encouragement of independent thought (by the teacher and the course)</td>
<td>.25 (13)</td>
<td>.33 (4)</td>
</tr>
<tr>
<td>17. Personality characteristics (&quot;personality&quot;) of the teacher</td>
<td>.24 (NA)</td>
<td>NA</td>
</tr>
<tr>
<td>18. Teacher's concern and respect for students; friendliness of teacher</td>
<td>.23 (14.5)</td>
<td>.65 (12)</td>
</tr>
<tr>
<td>19. Nature, quality, and frequency of feedback from teacher to students</td>
<td>.23 (14.5)</td>
<td>.87 (17)</td>
</tr>
<tr>
<td>20. Pleasantness of classroom atmosphere</td>
<td>.23 (NA)</td>
<td>NA</td>
</tr>
<tr>
<td>21. Nature and value of course material (including its usefulness and relevance)</td>
<td>.17 (16)</td>
<td>.70 (13)</td>
</tr>
<tr>
<td>22. Nature and usefulness of supplementary materials and teaching aids</td>
<td>-.11 (17)</td>
<td>.72 (14.5)</td>
</tr>
<tr>
<td>23. Difficulty of the course (and workload)—description</td>
<td>.09 (NA)</td>
<td>NA</td>
</tr>
<tr>
<td>24. Difficulty of the course (and workload)—evaluation</td>
<td>.07 (NA)</td>
<td>NA</td>
</tr>
</tbody>
</table>

Note: This table has been constructed using data from Tables 1 and 3 in Feldman, 1989a.
3) and "student-perceived outcome or impact of the course" (dimension 4) are the next most highly correlated with student achievement. Somewhat more moderate-sized correlations were found for the instructional dimensions from 5 through 11. Less strongly associated with student achievement are the dimensions from 12 through 21. Dimensions 22 through 24 are not related to student achievement, and nothing could be concluded about dimensions 25 through 28.

**Relation Between Specific Aspects of Teaching and Overall Evaluation.** If each student's overall evaluation of an instructor is made up of the student's evaluation of specific aspects of the teacher and the instruction, weighted by the student's estimation of the relative importance of these aspects to good teaching, then the overall assessment of teachers would be more highly correlated with characteristics those students consider important to good teaching than with those considered less important (compare Crittenden and Norr, 1973). To establish the differential importance of a dimension of teaching, I computed the correlation of each dimension with the global assessment of teachers. (See Feldman, 1976b, 1988, for limitations to this approach.) For each of the instructional dimensions (see Feldman, 1976b, Table 1 and fn. 5), standardized ranks were averaged across a number of pertinent studies. These averages are given in column 2 of Table 6.1 along with their rank from 1 to 17 in the parentheses.

The two analyses, each determining the importance of instructional dimensions from separate studies, have eighteen dimensions in common. The correlation between the ranks shown in parentheses for the separate studies is +.61. Thus, the instructional dimensions that are most highly associated with student achievement (column 1) also tend to be those that best discriminate among teachers with respect to the overall evaluation they receive from students (column 2). Note, however, some discrepancies, especially for "teacher's availability and helpfulness" (high when correlated with achievement but low when correlated with global evaluations) and for "intellectual challenge and encouragement of independent thought" (low when correlated with achievement but high when correlated with global evaluations).

If, relative to the other dimensions, ranks 1 through 6 are thought of as dimensions of high importance, ranks 7 through 12 of moderate importance, and ranks 13 through 17 of low importance, then the results of the two methods show the following pattern.

Both methods indicate that the teacher's preparation, course organization, teacher clarity and understandableness, teacher's stimulation of students' interest, and student-perceived outcome or impact are of high importance. (Items ranked in the top six in both studies.) Teacher's elocutionary skill, clarity of course objectives and requirements, teacher's knowledge of subject and enthusiasm are of moderate importance relative to other dimensions. (Items ranked between 7 and 12 on both measures.)
Nature, quality, and frequency of feedback to students; nature and value of course material; and nature and usefulness of supplementary materials and teaching aids were of low importance. (Items ranked 13 or lower on both measures.)

Causal Relationships and Other Considerations. Returning to the meaning of the correlations found between specific dimensions and student achievement (column 1), it is important to recognize that these correlations, by themselves, do not establish causal connections between teacher behavior and student achievement. Rather, as Leventhal (1975) points out, some third variable such as student motivation, ability, or aptitude might independently affect both teacher performance and student learning. Establishment of causal connections would require truer experimental designs and the control of extraneous variables. However, where such precautions were taken (Cohen, 1980b, Feldman, 1989a), results were similar to studies where these variables could not be controlled.

Experimental research on cause-effect relationships has been reviewed (selectively) by Murray (1991) who 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). This supports the findings reported in Table 6.1.

Enthusiastic and expressive attitudes and behaviors of teachers are highlighted in Murray’s analysis. In the correlations in Table 6.1, the instructional dimensions of “teachers’ enthusiasm (for subject or for teaching)” and “teacher’s elocutionary skills” (assumedly an aspect of enthusiasm and expressiveness) are also associated with achievement but only moderately so compared to some of the other instructional dimensions. Note, however, 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); and, interestingly enough, as the present discussion has emphasized, the instructional dimension of “teacher’s stimulation of interest in the course and its subject matter” has been found to be rather highly correlated with students’ achievement (ranked fourth) and highly associated, as well, with global evaluation of instruction relative to the other instructional dimensions (ranked first), thus lending support to Murray’s observation. Both Murray’s analysis and the present analysis emphasize the importance of teacher clarity for effective instruction.

Underlying Mechanisms and Other Considerations. Whether the associations between student learning and teacher’s attitudes, behaviors, and practices are established by correlational or experimental studies, we need to know more about the exact psychological and social psychological mechanisms by which these characteristics influence learning. When a large association between an instructional characteristic and student achievement is found, the tendency is to see the finding as self-explanatory. 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.

In fact, however, we need to specify exactly how those dimensions that affect achievement do so. Indeed, conceptually and empirically specifying such mechanisms in perhaps the most "obvious" connection of them all—that between student achievement and clarity and understandableness of instructors—has turned out to be particularly complex. Likewise, the mechanisms underlying the correlation between the teacher's organization and student achievement have yet to be specifically and fully determined.

**Interpreting Teacher and Course Evaluations to Identify Exemplary Teachers**

I now turn to the topic of identifying the exemplary teachers. Here students' global, or overall, evaluations of their teachers (in addition to the more specific evaluations already discussed) are often used.

It is at this point that certain problems and issues in interpreting student ratings become particularly evident. One way of exploring these problems and issues is to focus on the half-truths, if not outright myths, that have cropped up about student assessment and that are, in part, responsible for the unease felt by some faculty, administrators, and students in using student evaluations to identify exemplary teachers.

**Myths about Student Evaluations.** Aleamoni (1987) has listed a number of "myths" about student ratings of instructors and instruction. I agree that, at least as far as current research shows, the following are untrue (evidence that research does not support these contentions can be found in the following reviews: Cohen, 1980a; 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):

Students cannot make consistent judgments about the instructor and instruction because of their immaturity, lack of experience, and capriciousness.

Only colleagues with excellent publication records and expertise are qualified to teach and to evaluate their peers' instruction—good instruction and good research are so closely allied that it is unnecessary to evaluate them separately.

Most student rating schemes are nothing more than a popularity contest, with warm, friendly, humorous instructors emerging as winners every time.

Students are not able to make accurate judgments until they have been away from the course and possibly from the university for several years.

Student ratings are both unreliable and invalid.

The time and day the course is offered affects student ratings.

Student ratings cannot be used meaningfully to improve instruction.

For the most part, Aleamoni (1987) also seems correct in calling the following statement a myth: "Gender of the student and the instructor..."
student ratings." Consistent evidence cannot be found that either male or female college students routinely give higher ratings to teachers (Feldman, 1977). As for gender of the teacher, a recent review (Feldman, 1993) showed that a majority of the relevant studies found male and female college teachers not to differ in global ratings. However, across studies, evidence suggests that students may rate same-gendered teachers a little more highly than they do opposite-gendered teachers, although other variables may contribute to this finding.

Half-Truths and the Question of Bias. Although Aleamoni (1987) presents the following statements as candidates for status of myth, in reality there are small associations between certain factors and student ratings, as follows:

The size of the class affects student ratings (slightly higher ratings are given to teachers of smaller rather than larger courses; see Feldman, 1984; Marsh, 1987).

The level of the course affects student ratings (slightly higher ratings are given to upper-level than lower-level courses; see Feldman, 1978).

The rank of the instructor affects student ratings (slightly higher ratings are given to teachers of higher rather than lower academic ranks; see Feldman, 1983; Marsh, 1987).

Whether students take the course as a requirement or an elective affects the ratings (slightly higher ratings are given to elective courses; see Feldman, 1978; Marsh, 1987).

Whether students are majors or nonmajors affects their ratings (slightly higher ratings are given by students who are majors; see Feldman, 1978; Marsh, 1987).

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 (for the conceptualization of "bias" used here, see Marsh, 1987). For 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 the ratings are biased by students' expressing their dislike of large classes by rating instructors a little lower than they otherwise would. To put the matter in more general terms, certain course characteristics and situational contexts may indeed affect teaching effectiveness, and student ratings may then accurately reflect differences in teaching effectiveness. However, those making decisions about promotions and teaching awards should consider that it may be somewhat harder to be effective in some courses than others.

The idea that "the grades or marks students receive in the course are highly correlated with their ratings of the course and instructor" (Aleamoni, 1987) is another candidate for the status of myth, since grades are not highly correlated with student ratings. However, almost all of the available research does show a small or even modest positive association between grades and
evaluation (see Feldman, 1976a, 1977; Stumpf and Freedman, 1979). Research has also shown that some part of the positive correlation has a basis in reality and therefore is unbiased: students who learn more earn higher grades and thus legitimately give higher evaluations.

The academic discipline of a course is yet another correlate of—and therefore, possible influence on—teacher and course evaluations. Reviewing eleven studies, I found that teachers in different academic fields tend to be rated somewhat differently, with teachers in English, humanities, arts, and languages being rated higher than those in the social sciences, followed by those in the natural sciences, mathematics and engineering (Feldman, 1978). More recent data using two national evaluation instruments (IDEA and SIR) reported differences similar to those in my review (Cashin, 1990; Cashin and Clegg, 1987; Cashin and Sixbury, 1993). Among possible causes of these differences are 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, motivations, learning styles, and perceptions of good teaching. However, Cashin and Sixbury (1993) caution institutions to look carefully at disciplinary differences in evaluations to be sure that they are a function of something other than teaching effectiveness before adjusting the results of students' evaluations in different disciplines.

Conclusion

Nothing I have written here is meant to imply that course and teacher evaluations are the only means of identifying exemplary teachers and teaching. Discussions about teaching portfolios (see, for example, Edgerton, Hutchings, and Quinlan, 1991, and Chapter Nine of this volume) emphasize the importance of diverse information from a variety of sources. Yet, when properly administered and interpreted, the global and specific items contained in student surveys can serve as an important source of information for identifying exemplary teachers and teaching.

References


---

KENNETH A. FELDMAN is professor of sociology at the State University of New York, Stony Brook.