Through a half-century of research on student ratings, the constant quest has been to prove or disprove the existence of biasing factors. What have we learned, and what has happened as a result?

Looking for Bias in All the Wrong Places: A Search for Truth or a Witch Hunt in Student Ratings of Instruction?

Michael Theall, Jennifer Franklin

Few issues in higher education are as sensitive, divisive, and political as faculty evaluation and in particular the quality and value of the information provided by students in their evaluations of teachers and courses. Here are three statements that typify the polarity and problems in this issue. The first is from one of the most extensive and widely cited reviews of research on ratings. The second is a response by Marilley (1998) to an article by Wilson (1998) in the *Chronicle of Higher Education*. The third is a comment by someone responsible for the administration of a ratings process at a university.

> Ratings are 1) multidimensional; 2) reliable and stable; 3) primarily a function of the instructor who teaches the course rather than the course that is taught; 4) relatively valid against a variety of indicators of effective teaching; 5) relatively unaffected by a variety of variables hypothesized as potential biases; and 6) seen to be useful by faculty, . . . students, . . . [and] administrators [Marsh, 1987, p. 255].

> New evidence must be found to overturn the view that evaluations reveal who really knows how to teach, or more accurately, who knows how to make learning fun [Marilley, 1998; emphasis added].

> I provide evaluation services for my own and other institutions and I have received many requests to present the data in certain ways. One department

> chair, wanting to rank faculty, asked me to produce reports of average scores to the third decimal point. No matter how good the data, assuming this level of precision is greatly overestimating the discriminating power of ratings and is grossly unfair to the faculty. The problem is not with the ratings but with their use. This is not a reason to do away with ratings. Rather it is a reason to improve understanding and overall practice [Jennifer Franklin].

As these quotes suggest, evidence exists to support the validity and reliability of ratings, but there is a strong current of opinion not only against ratings but also actively seeking contradictory evidence. One must wonder about the extent to which those who seek contradictory opinions will be willing to accept existing research, no matter how substantial and replicated it has been. Finally, the final comment points to perhaps a more important issue than the methodological and psychometric questions surrounding ratings research. It is that data can be and are misused on a regular basis. Even if ratings results were perfectly reliable and valid (and no educational, psychological, or sociological instrument provides data that are perfect), misuse would still be a major problem.

For all these reasons, student ratings of teaching are hotly debated. Unfortunately, these debates are often uninformed by the extensive research done on the topic. That research (for example, the extensive review by Herbert Marsh in 1987) tells us that student ratings are generally valid and reliable and that they can provide valuable information for students, teachers, and administrators. Even when the data are technically rigorous, one of the major problems is day-to-day practice: student ratings are often misinterpreted, misused, and not accompanied by other information that allows users to make sound decisions. When we (Franklin and Theall, 1989) surveyed several hundred faculty and administrators, we found a surprising lack of knowledge about the literature of student ratings and even about the basic statistical information necessary to interpret ratings reports accurately. That lack of knowledge correlated significantly with negative opinions about evaluation, student ratings, and the value of student feedback. We also surveyed faculty and staff in teaching centers or similar instructional support units—people with training and experience in the use of evaluation data. This group had significantly higher scores than the faculty-and-administrator group on the knowledge portion of the survey and had much more positive attitudes about students and the value of ratings information. The difference between the two groups is important taken in light of research, reviews, and applications works (for example, Cohen, 1980; McKeachie, 1987; Theall and Franklin, 1991) that have shown that when ratings information is coupled with knowledgeable assistance for formative purposes, improvement can result. Why, then, the resistance to ratings and the seemingly never-ending search for biases that might disprove their validity or value?

As the studies cited and the consensus of researchers and practitioners attests (Theall, 1994), one part of the answer lies in poor summative prac-

tice. The absence of clear policy, the use of poor instrumentation, the misuse and misinterpretation of data, and arbitrary decision making have all led to situations that contradict the literature (Theall, 1996a) and as a result, many faculty cite instances in which the established literature seems to have been disproved.

Another part of the answer lies in the psychological literature on topics such as efficacy (Bandura, 1977), attributions (Weiner, 1986), and expectancy (Jones, 1977). The notion of having someone else determine the quality of one's work is threatening, and when the evaluators of the work are not considered to be as qualified as the evaluatee, anxiety and resistance can increase. Boice (1992) documents the disenchantment of many new faculty who, despite conscientious efforts to prepare for their courses, still face student criticism. He concludes that many new faculty overprepare and concentrate so completely on the delivery of content that they exclude time for discussion, questions, dialogue, and other opportunities for interaction with students, a very important element in successful teaching and learning in and out of the classroom (Pascarella and Terenzini, 1991). Although the lower ratings accurately reflect student dissatisfaction, they can be inaccurately interpreted as meaning that the teacher is not doing an adequate job. The truth of the matter, as Boice (1992) has shown, is that some simple changes can result in increased ratings without sacrificing content or the quality of teaching and learning. However, in the face of these negative ratings and without instructional support, faculty who have prepared long and hard have to reconcile certain knowledge of effort expended against a lack of success. No wonder, then, that these faculty may develop negative attitudes toward students and student ratings. Even more serious, if the situation persists over time, a pathological pattern of behavior can develop that can lead to serious psychological problems. The stages of "professorial melancholia" (Machell, 1989) include increasing hostility toward students and administrators and, eventually, arrogance, alienation, and even possible substance abuse and verbal or grade abuse of students.

It is no wonder, then, that so much effort has been committed to seeking negative evidence. Unfortunately, most of this effort has been misdirected, trying to prove that the ratings data are biased, when it should have been directed at increasing the skills of users of the data and at correcting problems with day-to-day practice.

Reports of bias in ratings often get wide circulation (as in articles in *Change* magazine by Trout, 1997, and Williams and Ceci, 1997), but the truth is that the vast majority of these reports of invalidity or bias have been essentially refuted. Even the most widely discussed reports (for example, the "Dr. Fox" study by Naftulin, Ware, and Donnelley, 1973, and the report of negative correlations between ratings and learning by Rodin and Rodin, 1972) were unreplicable and were shown to be flawed in their conceptualization or execution. In a series of studies correcting the Dr. Fox flaws, Perry and associates (Perry, Abrami, and Leventhal, 1979; Perry, Magnusson, Parsonson,

and Dickens, 1986) demonstrated that while content was always critical to learning, improving presentational skill and style resulted in better overall ratings without sacrificing the quality of learning. The studies also showed that style alone was not a substitute for content and that students recognized the difference. In the most widely cited study of the relationship between ratings and learning, Cohen (1981) found significant correlations (> .40) between ratings and student performance on common final examinations in multisection classes. Having the exams corrected by someone other than the instructors of the sections avoided grading bias. This study and replications of it form the foundation for the ratings-learning relationship, and no evidence has yet surfaced to refute Cohen's findings.

Ratings Myths and Research Evidence

There are many misconceptions about student ratings of instruction. Several writers (for example, Aleamoni, 1987) have presented these issues, but the misconceptions persist. We shall discuss some of the most common myths about ratings and look at the evidence from research on these issues. Each issue is first presented as a question, and relevant research is then discussed.

Are Students Qualified to Rate Their Instructors and the Instruction They Receive? The myth says no, but generally speaking, the answer is yes. Part of the dispute centers on the definition of the term *qualified* and on the intent of the evaluation. Opponents of ratings (Trout, 1997) essentially state that students are not qualified to rate any aspect of teaching. Individuals who are more involved in the research and practice of evaluation (Arreola, 1994; Theall and Franklin, 1990a) disagree, noting that in some areas, students are well qualified.

Students spend a full term in the course, observe the instructor in class and in interactions with students, and can accurately judge what or how much they have learned with respect to their knowledge at entry. Students can report the frequencies of teacher behaviors, the amount of work required, and the difficulty of the material. They can answer questions about the clarity of lectures, the value of readings and assignments, the clarity of the instructor's explanations, the instructor's availability and helpfulness, and many other aspects of the teaching and learning process. No one else is as qualified to report on what transpired during the term simply because no one else is present for as much of the term. Peers and administrators can visit the class, but such visits usually occur only once or twice per term, and although such visits are valuable, they cannot come close to equaling the range of events on which students base their opinions. Peers and administrators are also generally more knowledgeable of the content and thus cannot necessarily empathize with the views of students who may be having problems. Because students have this long-term exposure, it is also reasonable to ask them to summarize their opinions in some overall ratings of the instructor and the course.

But students are not necessarily qualified to report on all issues. For example, beginning students do not have sufficient depth of understanding to accurately rate the instructor's knowledge of the subject. They might estimate knowledge based on the instructor's ability to respond to questions, but this estimate is probably less valuable than a colleague's rating if the purpose is to assess the depth and breadth of the instructor's knowledge. Students are certainly qualified to express their satisfaction or dissatisfaction with the experience. They have a right to express their opinions in any case, but no one else can report the extent to which the experience was useful, productive, informative, satisfying, or worthwhile. While opinions on these matters are not direct measures of the performance of the teacher, they are legitimate indicators of student satisfaction, and there is a substantial research base linking this satisfaction to effective teaching. There is also a logical and undeniably pragmatic reason to attend to student views. They enroll in classes and pay tuition. Higher education can no longer afford to take an elitist approach that dismisses all but those who agree with its policies or procedures and who sit in silent awe at the feet of those who "profess."

Are Ratings Based Solely on "Popularity"? The myth here is that a popular teacher is not a good teacher. There is no basis for this argument and no research to substantiate it. When this myth is brought out, the term *popular* is never defined. Rather, it is left to imply that learning should somehow be unpleasant, and the "popularity" statement is usually accompanied by an anecdote suggesting that the best teachers are the ones students dislike the most. Theall (1998) reviewed comments made in reaction to a January 1998 article on student ratings that appeared in the *Chronicle of Higher Education* (Wilson, 1998) and provided many examples of such unsubstantiated claims. The assumption that popularity somehow means a lack of substance or knowledge or challenge is totally without merit. There are no studies to support this view.

Are Ratings Related to Learning? The most acceptable criterion for good teaching is student learning. There are consistently high correlations between students' ratings of the "amount learned" in the course and their overall ratings of the teacher and the course. Even more telling, in studies in multisection courses that employ a common final exam, the students who gave the highest ratings to their instructors were the ones who performed best on their exams (Cohen, 1981). Those who learned more gave their teachers higher ratings. These studies are the strongest evidence for the validity of student ratings because they connect ratings with learning.

Can Students Make Accurate Judgments While Still in Class or in School? The myth says that students can discern real quality only after years of experience in the workforce. There is no research proving this statement, but there have been several studies comparing ratings in class to ratings by the same students in the next term, the next year, immediately after graduation, and several years later (for example, Centra, 1979; Frey, 1976). There have also been studies of instructor performance over time (Marsh, 1992) showing consistent ratings of teachers by students over periods as

long as thirteen years. All these studies report the same results: although students may realize later that a particular subject was more important that they thought, student opinions about teachers change very little over time. Teachers rated highly in class are rated highly later on, and those with poor ratings in class continue to get poor ratings later on. Teachers rated highly by one group tend to be rated highly by other groups.

Are Student Ratings Reliable? This is more a technical question. The myth says no; the research says yes. Whether reliability is measured within classes, across classes, over time, or in other ways, student ratings are remarkably consistent. Marsh's review (1987) provides the most comprehensive array of evidence supporting this view.

Does Gender Make a Difference? Reviews of gender studies (Centra and Gaubatz, 1998; Feldman, 1992a, 199b) have reached similar conclusions: there is no strong or regular pattern of gender-based bias in ratings. That is, students do not favor instructors on the basis of gender alone. There are a few studies that suggest other kinds of gender bias in higher education. For example, one study (Franklin and Theall, 1992) found that female instructors in one department were largely assigned entry-level, required, large-enrollment courses while males disproportionately taught upper-level and graduate seminars. Considering that certain research indicates that ratings in the first group of courses will be a bit lower, such course assignments automatically put the female instructors at risk. Further, if interpretation of ratings results simply arrayed average scores by gender, females would have lower scores. The result would be an incorrect and unfair evaluation of the female faculty. The scores would reflect the differences in teaching situations but *not* that female instructors were less competent and *not* that students were biased against female faculty.

Are Ratings Affected by Situational Variables? The research says that ratings are robust and not greatly affected by such variables (Marsh, 1987). But we must keep in mind that generalizations are not absolute. There will always be variations. For example, we know that required, large-enrollment, out-of-major courses in the physical sciences get lower average ratings than elective, upper-level, in-major courses in literally all disciplines. Does this mean that teaching quality varies? Not necessarily. What it does show is that effective teaching and learning may be harder to achieve under certain sets of conditions. The saving grace here is that the overall effect of such variables is small.

Do Students Rate Teachers on the Basis of Expected or Given Grades? This is currently the most contentious question in ratings research. There is consistent evidence of a relationship between grades and ratings: a modest correlation of about .20. The multisection validity studies (for example, Cohen, 1981) consider this relationship to be an expected phenomenon because it follows from a learning-satisfaction relationship. The multisection studies, with their correlations above .40, still provide the most solid evidence that ratings reflect learning. These findings lead

to the conclusion reached by most researchers that there should be a relationship between ratings and grades because good teaching leads to learning, which leads to student achievement and satisfaction, and ratings simply reflect this sequence. Recent and rigorous studies by Greenwald and Gillmore (1997) claim that all else being controlled, giving higher grades ("grade inflation") can raise ratings. In a debate on the issues held at the annual meeting of the American Educational Research Association, Abrami and d'Apollonia (1998) and Marsh and Roche(1998) debated Greenwald and Gillmore's contentions, questioning the research and arguing that the presence of a grades-ratings relationship does not refute the established connection between ratings and learning. The question at this point becomes an ethical one: "Is giving higher grades in order to get higher ratings a problem with ratings or a problem with ethics, and should attempts to correct the problem be psychometric or policy issues?

Basic Considerations for Good Evaluation Practice

One of the first issues in evaluation is to determine its purpose. When we gather information to review or explore or improve, we describe this as "formative evaluation." When our purpose is to make decisions about merit, promotion, or tenure, for example, we call it "summative evaluation." Theall and Franklin (1990b) point out the need to consider a complex matrix of purposes, sources, and users in any summative evaluation, particularly when teaching performance is being assessed. Though it may seem obvious that summative evaluation includes more technical rigor and a wider array of data, the unfortunate reality is that summative decisions about teaching are often made on the basis of student ratings data alone. As a result, there is a great deal of suspicion, anxiety, and even hostility toward ratings.

Evaluation is a systematic process and requires acceptance, participation, and cooperation from a number of stakeholders. There are ways to develop evaluation systems that take into account the complexity and sensitivity of the process. As Arreola (1994) demonstrates, arriving at consensus about what is important, what will be evaluated, who will contribute, and what criteria will be used is the most important first step in good practice.

Student ratings are only one source of information about teaching, and teaching is only one aspect of faculty performance. Never make the mistake of judging teaching or overall performance on the basis of ratings alone. Research on student ratings has given us consistent findings, and Marsh (1987) has outlined these as definitively as anyone. But research findings generalize from a sample to a population and do not guarantee that every situation will be explained. It is critical to have an understanding of the context of the evaluation so as to be able to make fair and accurate decisions. To be fair, comparisons of faculty teaching based on ratings should use sufficient amounts of data from similar situations. It would be grossly unfair to compare the ratings of someone teaching a graduate seminar with ten

students to the onetime ratings of someone teaching an entry-level required course with an enrollment of two hundred. Common sense, research, and ethical practice all demand correct interpretation and use of evaluation data.

Here is a set of guidelines for good evaluation practice.

- *Establish the purpose of the evaluation and the uses and users of ratings beforehand.* Do this by including all who will be involved in or affected by the process. Identify what is important and what should and will be evaluated, and go on to establish what kinds of data will be collected, who will provide the data, how they will be analyzed, whether all data will have equal weight, how the data will be assembled for users, and how data will be used in decision making.
- *Include all stakeholders in decisions about evaluation process and policy.* As indicated in this chapter and in the literature (for example, Arreola, 1994; Centra, 1979; Miller, 1987), developing evaluation policy or process in the absence of the individuals who will be affected is a serious error. Including sufficient time to involve the stakeholders and carrying out processes establishing consensus are part of this consideration.
- *Publicly present clear information about the evaluation criteria, process, and procedures.* As part of the a priori decision-making process and after such decisions are made, aggressively publicize the intent, purposes, and process of evaluation, emphasizing its potential to support improvement.
- *Produce reports that can be understood easily and accurately.* No matter how well designed the instrumentation, the evaluation system will face problems if the reports it generates are overly complicated, are difficult to interpret, or present data in ways conflicting with the purposes of the evaluation. Formative reports should be detailed and coupled with information and advice about the meaning and implications of the data. If improvements are needed, some suggestions for action are important. Summative reports should be clear, unambiguous, and more general and should allow users to make necessary decisions based on agreed-to arrays of data that employ accepted norms or decision criteria. Summative reports should also contain information important to understanding the context of the evaluation (for example, ratio of students enrolled in the class to those responding to the evaluation; level of the course; required versus elective status; some student demographics). Graphic displays using confidence intervals clearly showing when individuals significantly differ from comparison groups can help users avoid misinterpretation. We have identified several factors important to the design of useful reports. (see Franklin and Theall, 1990).
- *Educate the users of ratings results to avoid misuse and misinterpretation.* Particularly critical to effective evaluation is maintaining an ongoing cycle of training emphasizing the correct interpretation and appropriate use of the evaluation data. Given widespread misuse of data and misunderstanding about its interpretation, this can be the most important aspect of day-to-day practice.

- *Keep a balance between individual and institutional needs in mind.* Evaluation can and should serve both institutional and individual needs. It is possible to create complete systems for both evaluation and development, and such systems benefit faculty, students, and institutions because they ultimately support better teaching and learning.
- *Include resources for improvement and support of teaching and teachers.* This is part of a complete system and cannot be omitted. Evaluation without it is punitive. Evaluation accompanied by visible and effective development becomes a valued component of teaching and learning and the process of personnel decision making. One of the major factors in creating a campus culture and climate that support teaching is to have an established center for teaching and qualified staff to provide assistance in instructional design, development, and evaluation. Research shows us that teachers benefit most from evaluation data when the data are competently explained and when assistance and resources for improvement are available. Simply sending a computer printout to a teacher does little to help that teacher understand the results or to improve teaching. Commitment to and support for teaching from the highest levels of the institution are required if the evaluation process it to be perceived as useful and nonthreatening. Anything less results in polarized views about the purpose of evaluation and leads to anxiety, resistance, and hostility.
- *Keep formative evaluation confidential and separate from summative decision making.* Even though it is possible to develop a comprehensive system that serves formative and summative purposes, it is critical to separate the two purposes conceptually and in practice. Establish policy guidelines for the distribution and use of data, and get the commitment and active support of faculty and administrators for adherence to these policies. Allow formative evaluation to explore innovative techniques without the threat of failure. Use formative data for classroom assessment and research, but do not make personnel or program decisions without agreement about what kinds of data are appropriate and how such data should be used.
- *Adhere to rigorous psychomteric and measurement principles and practices.* Use, adapt, or develop instrumentation specific to the purposes and needs of the situation. Maintain databases, and validate instruments before using data summatively. Conduct data analysis regularly to establish norms or criteria and to clarify institutional differences across departments, disciplines, or demographic groups. Revise interpretation guidelines on the basis of clear analysis and understanding of the data. Bring in independent outside experts, if necessary, to assist in the development and validation of instruments and processes.
- *Regularly evaluate the evaluation system.* Conditions change, and the evaluation system must change to adapt to new conditions. Regular evaluation of the performance of the evaluation system is necessary to ensure that it is accurate, timely, efficient, and effective and that policies and processes

are appropriate and being adhered to. When institutional or programmatic changes are made, review the evaluation system and adapt it as needed. Again, seek expert advice and assistance when necessary.

- *Establish a legally defensible process and a system for grievances.* Miller (1987) rightly establishes this as an important issue. Without guarantees of protection from mistakes or misuse, faculty and the institution are at risk. The best insurance against unpleasant surprises in this area is the work done before any data are ever collected, that is, the consensus and public exploration processes discussed in the first three items in this list.
- *Consider the appropriate combination of evaluation data with assessment and institutional research information.* Evaluation data can shed light on program and school performance and can provide important information for purposes of assessment and even accreditation. Classroom research and assessment can be supported and institutional questions can be addressed as well. The synergy of the resources devoted to evaluation, assessment, and institutional research has tremendous potential. Combining these complementary but often isolated efforts can result in better understanding of overall institutional performance, of student learning and satisfaction, of teaching and learning issues, and of other matters of importance to all members of a higher education community as well as to other constituencies such as legislators, trustees, and boards of higher education. The opportunity to take advantage of this potential should not be overlooked.

Conclusion

The principal themes of this chapter are few and straightforward. First, student ratings and other evaluation data can provide powerful and useful information. Second, good evaluation practices and the attendant benefits must be based on a systematic and careful approach involving all constituencies and achieving consensus on major issues. Third, appropriate and accurate interpretation and use of data is as important as rigorous statistical and analytical procedures. Finally, evaluation must be appropriately supported and coupled with equivalent support for improvement, recognition, and rewards. The issues, the decisions, and the future are too important to allow haphazard processes, inaccurate data, and misuse of results. Faculty, students, institutions, and higher education itself require and will benefit most from comprehensive systems of evaluation and the synergy of institutional efforts to identify and promote excellence.

References

Abrami, P. C., and d'Apollonia, S. "The Positive Relationship Between Course Grades and Course Ratings: What Is the Cause and What, If Anything, Can Be Done About It?" Debate presented at the 79th Annual Meeting of the American Educational Research Association, San Diego, Apr. 1998.

Aleamoni, L. M. "Typical Faculty Concerns About Student Evaluation of Teaching." in L. M. Aleamoni (ed.), *Techniques for Evaluating and Improving Instruction.* New Directions for Teaching and Learning, no. 31. San Francisco: Jossey Bass, 1987.

Arreola, R. A. *Developing a Comprehensive Faculty Evaluation System.* Boston: Anker, 1994.

Bandura, A. "Self-Efficacy: Toward a Unifying Theory of Behavioral Change." *Psychological Review,* 1977, *84,* 191–215.

Boice, R. *The New Faculty Member: Supporting and Fostering Professional Development.* San Francisco: Jossey Bass, 1992.

Centra, J. A. *Determining Faculty Effectiveness.* San Francisco: Jossey Bass, 1979.

Centra, J. A., and Gaubatz, N. B. "Is There Gender Bias in Student Evaluations of Teaching?" Paper presented at the 79th Annual Meeting of the American Educational Research Association, San Diego, Apr. 1998.

Cohen, P. A. "Effectiveness of Student-Rating Feedback for Improving College Instruction: A Meta-Analysis." *Research in Higher Education,* 1980, *13,* 321–341.

Cohen, P. A. "Student Ratings of Instruction and Student Achievement: A Meta-Analysis of Multisection Validity Studies." *Review of Educational Research,* 1981, *51,* 281–309.

Feldman, K. A. "College Students' Views of Male and Female College Teachers, Part 1: Evidence from the Social Laboratory and Experiments." *Research in Higher Education,* 1992a, *33,* 317–375.

Feldman, K. A. "College Students' Views of Male and Female College Teachers, Part 2: Evidence from Students' Evaluations of Their Classroom Teachers." *Research in Higher Education,* 1992b, *33,* 415–474.

Franklin, J. L., and Theall, M. "Who Reads Ratings: Knowledge, Attitudes, and Practice of Users of Student Ratings of Instruction." Paper presented at the 70th Annual Meeting of the American Educational Research Association, San Francisco, Mar. 1989.

Franklin, J. L., and Theall, M. "Communicating Ratings Results to Decision Makers: Design for Good Practice." In M. Theall and J. L. Franklin (eds.), *Student Ratings of Instruction: Issues for Improving Practice.* New Directions for Teaching and Learning, no. 43. San Francisco: Jossey Bass, 1990.

Franklin, J. L., and Theall, M. "Student Ratings of Instruction and Gender Differences Revisited." Paper presented at the 75th Annual Meeting of the American Educational Research Association, New Orleans, Apr. 1992.

Frey, P. W. "Validity of Student Instructional Ratings. Does Timing Matter?" *Journal of Higher Education,* 1976, *3,* 327–336.

Greenwald, A. G., and Gillmore, G. M. "Grading Leniency Is a Removable Contaminant of Student Ratings." *American Psychologist,* 1997, *52,* 1209–1217.

Jones, R. A. *Self-Fulfilling Prophecies: Social, Psychological, and Physiological Effects of Expectancies.* New York: Halsted Press, 1977.

Machell, D. F. "A Discourse on Professorial Melancholia." *Community Review,* 1989, *9*(1–2), 41–50.

Marilley, S. M. "Response to 'Colloquy.' " *Chronicle of Higher Education.* [http://chronicle.com/colloquy/98/evaluation/09.htm]. 1998.

Marsh, H. W. "Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research." *International Journal of Educational Research,* 1987, *11,* 253–388.

Marsh, H. W. "A Longitudinal Perspective of Students Evaluations of University Teaching: Ratings of the Same Teachers over a Thirteen-Year Period." Paper presented at the 73rd Annual Meeting of the American Educational Research Association, San Francisco, Apr. 1992.

Marsh, H. W., and Roche, L. A. "Effects of Grading Leniency and Low Workloads on Students' Evaluations of Teaching." Paper presented at the 79th Annual Meeting of the American Educational Research Association, San Diego, Calif., Apr. 1998.

McKeachie, W. J. "Can Evaluating Instruction Improve Teaching?" In L. M. Aleamoni (ed.), *Techniques for Evaluating and Improving Instruction*. New Directions for Teaching and Learning, no. 31. San Francisco: Jossey-Bass, 1987.

Miller, R. I. *Evaluating Faculty for Promotion and Tenure*. San Francisco: Jossey-Bass, 1987.

Naftulin, D. H., Ware, J. E., and Donnelly, F. A. "The Doctor Fox Lecture: A Paradigm of Educational Seduction." *Journal of Medical Education*, 1973, *48*, 630–635.

Pascarella, E. T., and Terenzini, P. T. *How College Affects Students: Findings and Insights from Twenty Years of Research*. San Francisco: Jossey-Bass, 1991.

Perry, R. P., Abrami, P. C., and Leventhal, L. "Educational Seduction: The Effect of Instructor Expressiveness and Lecture Content on Student Ratings and Achievement." *Journal of Educational Psychology*, 1979, *71*, 107–116.

Perry, R. P., Magnusson, J. L., Parsonson, K. L., and Dickens, W. J. "Perceived Control in the College Classroom: Limitations in Instructor Expressiveness Due to Noncontingent Feedback and Lecture Content." *Journal of Educational Psychology*, 1986, *78*, 96–107.

Rodin, M., and Rodin, B. "Student Evaluations of Teachers." *Science*, 1972, *177*, 1164–1166.

Theall, M. "What's Wrong with Faculty Evaluation: A Debate on the State of the Practice." *Instructional Evaluation and Faculty Development*, 1994, *14*(1–2), 27–34.

Theall, M. "When Meta-Analysis Isn't Enough: A Report of a Symposium About Student Ratings, Conflicting Results, and Issues That Won't Go Away." *Instructional Evaluation and Faculty Development*, 1996a, *15*(1–2), 1–14.

Theall, M. "Who is Norm and What Does He Have to Do With Student Ratings?: A Reaction to McKeachie." *Instructional Evaluation and Faculty Development*, 1996b, *16*(1), 7–9.

Theall, M. "Colloquy, Colloquia, Colloquiarum: A Declining Form, a Questionable Forum." *Instructional Evaluation and Faculty Development*, 1998, *18*(1–2) http://www.uis.edu/ctl/sigfted.html.

Theall, M., and Franklin, J. L. "Student Ratings in the Context of Complex Evaluation Systems." In M. Theall and J. L. Franklin (eds.), *Student Ratings of Instruction: Issues for Improving Practice*. New Directions for Teaching and Learning, no. 43. San Francisco: Jossey-Bass, 1990a.

Theall, M., and Franklin, J. L. (eds.). *Student Ratings of Instruction: Issues for Improving Practice*. New Directions for Teaching and Learning, no. 43. San Francisco: Jossey-Bass, 1990b.

Theall, M., and Franklin, J. L. "Using Student Ratings for Teaching Improvement." In M. Theall and J. L. Franklin (eds.), *Effective Practices for Improving Teaching*. New Directions for Teaching and Learning, no. 48. San Francisco: Jossey Bass, 1991.

Trout, P. A. "What the Numbers Mean: Providing a Context for Numerical Student Evaluations of Courses." *Change*, 1997, *29*(5), 24–30.

Weiner, B. *An Attributional Theory of Motivation*. New York: Springer-Verlag, 1986.

Williams, W. M., and Ceci, S. J. "How'm I Doing? Problems with Student Ratings of Instructors and Courses." *Change*, 1997, *29*(5), 13–23.

Wilson, R. "New Research Casts Doubt on Value of Student Evaluations of Professors." *Chronicle of Higher Education*, Jan. 16, 1998, pp. A1, A16.

MICHAEL THEALL is associate professor of educational administration and director of the Center for Teaching and Learning at the University of Illinois at Springfield.

JENNIFER FRANKLIN is director of the Center for Teaching and Learning at California State University, Dominguez Hills.