

What Do Students (or Professors) Know About Teaching?¹

J. Robert Warmbrod
Distinguished University Professor Emeritus
The Ohio State University

Those of us who earned undergraduate and graduate degrees prior to the late 1960s probably rarely, if ever, were given the opportunity to provide written evaluations of professors or courses. Since the 1970s, students' evaluations of teaching have increasingly become commonplace in American colleges and universities (Seldin, 1999). Students' evaluations are used variously for three purposes: as diagnostic feedback to faculty about their teaching; to provide information to students for selecting courses and instructors; and, as a measure of teaching effectiveness used by faculty committees and administrators in making personnel decisions, namely, decisions about tenure, promotion, and merit pay. Presently, it is almost universal in American colleges and universities to use student ratings of teaching as part of faculty personnel reviews (A. Kalash, Director, Center for Faculty and TA Development, The Ohio State University, personal communication, November 30, 2005).

The paper will focus exclusively on the use of student ratings of teaching effectiveness as one source of evidence in making personnel decisions. I will concentrate on the results of empirical research primarily conducted since 1970 pertaining to student ratings of instructional effectiveness. A writer in the *Chronicle of Higher Education* claimed that nearly 2,000 studies have been completed on the topic making it "the most extensive area of research in higher education" (Wilson, 1998). I will describe some aspects of Ohio State's policy and practice regarding the use of student ratings of teaching as an example of how universities use student ratings data.

It is not surprising that there is sharp disagreement within the academic community about the use of student ratings as a measure of an instructor's performance. One position is the conclusion of a prominent researcher and writer on the evaluation of teaching who says "student ratings tend to be statistically reliable, valid, and relatively free from bias . . . , probably more so than any other data used for evaluation" (Cashin, 1995). A faculty critic argues that student evaluations of faculty "infringe on instructional responsibilities of faculty by providing a control mechanism over curricular, course content, grading, and teaching methodology," thereby posing "a serious unrecognized infringement on academic freedom" (Haskell, 1997); another professor asserts that student ratings "are not just invalid and unrealistic; they are pernicious" (Trout, 1997). The Executive Editor of *Change*, a highly regarded journal of higher education, contends that student ratings data are misused by promotion and tenure committees and calls for a reexamination of the "assumptions behind a practice gone stale" (Marchese, 1997).

The Faculty Rules of The Ohio State University specifically require departments and schools to assure that students are given the opportunity to evaluate every course every time it is taught. A form, the Student Evaluation of Instruction – or SEI as it is known – is made available for use university wide; however, departments are free to design their own instrument.

¹ Paper presented to Torch Club of Columbus, Ohio, January 12, 2006.

Furthermore, the promotion and tenure guidelines at OSU require that a faculty member's documentation include evidence of students' evaluations of teaching for "every course taught in the past 5 years or since the date of hire, if less than 5 years ago."

What is the research evidence bearing on the pros and cons of the use of student ratings of instruction, particularly when the information yielded by these surveys is used by promotion and tenure committees and administrators? The examination of the issues will be organized around four questions.

Are student ratings of teaching effectiveness reliable?

Are student ratings valid?

What characteristics of instructors and courses have the potential to bias ratings?

What is the relationship between professors' teaching effectiveness and their research productivity?

Dimensions of Teaching

Researchers who specialize in the evaluation of college teaching have established that teaching is multidimensional, that is, teaching involves a number of behaviors and activities. It follows, then, that students' evaluations should record students' perceptions about the various dimensions of an instructor's teaching behavior and performance (Marsh & Dunkin, 1997). One model describes teaching in terms of three factors whereby the instructor is, first, a communicator demonstrated by elocutionary skills, clarity and understandableness, structure and organization, and enthusiasm; second, a facilitator indicated by extent of intellectual challenge, class discussion, and availability and helpfulness to students; and third, a manager demonstrated by classroom management, clarity of objectives, fairness and impartiality, feedback to students, and difficulty and workload of the course.

A critically important dimension of teaching is an instructor's knowledge of the subject being taught. There is agreement that an instructor's subject matter mastery, the extent to which course content is current and germane to the objectives of the course, the extent the course reflects a current and appropriate research base, as well as the academic rigor of the course are dimensions that peers, not students, are most capable of making valid evaluations.

OSU's Student Evaluation of Instruction form -- the SEI -- includes nine items pertaining to various dimensions of teaching. Students respond to each item using a five-point scale ranging from "Agree Strongly" to "Disagree Strongly." Three items pertain to the organization of the course and the instructor's role as a communicator; three items pertain to the instructor's rapport with students; and, three items pertain to teaching and learning issues. The tenth item asks students to rate the instructor on a five-point scale ranging from "Excellent" to "Poor." Students' responses to the 10 items are quantified on a five-point scale where 5.0 indicates the most favorable rating. The forms, completed anonymously, are scanned and scored electronically with the instructor provided a report for each course indicating the percentage of students enrolled in the course who responded and, for each item, the instructor's mean and standard deviation -- based on the five-point scale -- with comparable statistics for each item for a sample of courses with similar characteristics for the instructor's department and college and for the University.

Reliability

Are student ratings of teaching effectiveness reliable? In the measurement of educational and psychological phenomena, reliability refers to the consistency and stability of the scores yielded by the measuring instrument. Reliability is assessed by determining the extent of agreement among the students in a class in their responses to items on the instrument. The general rule is the higher the number of raters the more reliable the quantitative values yielded by the rating instrument. Measurement specialists have established that the reliability of class-average numerical ratings from a sufficient number of students in any one class – 10 or more students is the recommendation – compares favorably with the reliability of the best objective tests (Marsh & Dunkin, 1977) – technically, this indicates a reliability coefficient of .70 or higher where coefficient values approaching 1.0 indicate reliable, that is, consistent and stable measurement.

Validity

Are students' evaluations valid measures of teaching effectiveness? Validity of measurement poses the question whether the numerical scores yielded by student rating instruments actually measure the construct "teaching effectiveness." The primary criterion of the effectiveness of instruction is students' level of achievement indicated by scores on examinations and grades earned. Therefore, studies seeking to establish the validity of student ratings are designed to describe the relationship between student ratings of teachers and courses and students' level of achievement. A substantial positive relationship between student ratings and achievement is evidence that students' evaluations are valid indicators of "teaching effectiveness."

A multi-section validity paradigm is used by researchers to examine the ratings-achievement relationship. The paradigm requires courses with multiple sections with each section taught by a different instructor. Instructors of all sections teach to accomplish the same course objectives and use the same syllabus and textbook. Students in all sections complete a common final exam and assess anonymously the teaching effectiveness of their instructor using a standardized evaluation instrument. The statistic indicating the validity of student ratings of teaching effectiveness is the correlation coefficient describing the relationship between section-average student ratings and section-average final exam performance.

An early study examining the student ratings-achievement relationship that drew widespread attention was published in the journal *Science* in 1972. The study, conducted by professors in two California universities involved graduate teaching assistants for 12 sections of an undergraduate calculus course. The teaching assistants met with students in recitation sessions two hours each week. The professor in charge of the course – one of the authors of the article – lectured three days each week. At the end of the quarter students evaluated the teaching effectiveness of their teaching assistant – not the professor who delivered three hours of lecture each week – by responding to one question: "What grade would you assign to his (teaching assistant) total teaching performance?" Students' level of achievement was quantified as the grade students earned in the course. Section-mean student ratings for the 12 sections were correlated with the section-mean course grades resulting in a coefficient of -.75. The authors

concluded “that students are less than perfect judges of teaching effectiveness” and “good teaching is not validly measured by student evaluations.” (Rodin & Rodin, 1972). The study has been cited widely (e.g., Sheehan, 1975) as evidence of the invalidity of student ratings for making personnel decisions (Cohen, 1990).

One year after the California professors’ study was published, two additional articles on the validity of student ratings were published in *Science*. The authors of both articles, explicitly noting that their research did not exhibit the methodological deficiencies of the previously published study, reported substantial positive correlations between students’ evaluations and achievement.

A Northwestern University professor (Frey, 1973) also studied the validity of student ratings of teaching for students enrolled in multi-section calculus courses. Each section was taught by a regular faculty member using a common syllabus and text; students completed a common final exam. Students completed an instructional rating form that measured students’ perceptions about six dimensions of their course and classroom experience. Section-mean scores for the six teaching dimensions were correlated with section-mean final exam scores, resulting in positive correlation coefficients of .87 for students’ estimates of how much they had learned in the course, .75 for teacher presentation/communication, .69 for grading, .62 for organization/planning, .44 for workload, and .31 for teacher accessibility.²

A professor of pharmacology (Gessner, 1973) at the State University of New York at Buffalo studied the ratings-achievement relationship for 23 subject areas of a basic science course completed by sophomore medical students. Each subject area was taught by a faculty member who had sole responsibility for instruction. The achievement criterion was performance on the National Medical Board Examination. Students rated teaching effectiveness anonymously with regard to two dimensions of teaching; correlation coefficients describing the relationship between students’ achievement and their ratings were .77 for content/organization and .69 for presentation.

In 1981, a psychology professor at Dartmouth (Cohen, 1981) synthesized research investigating the relationship between student ratings of instruction and student achievement for 68 multi-section courses. He reported average correlations of .47 for “overall course rating” and .43 for “overall instructor rating.” Using the same data set to examine the relationship between achievement and student ratings for specific instructional dimensions, a sociology professor at the State University of New York at Stony Brook (Feldman, 1989a), reported validity coefficients of .32 or higher for all dimensions of teaching that were investigated.

A review of the published multi-section validity studies by researchers at Concordia University, Montreal (d’Apollonia & Abrami, 1997), resulted in the conclusion that student ratings of instructional effectiveness are consistently valid across different students, courses, and settings. Other researchers who have analyzed the multi-section validity research have concluded the analyses demonstrate that student ratings are valid indicators of teaching effectiveness (Marsh & Dunkin, 1997). Also, studies indicate that students’ evaluations of instruction are

² In social sciences, validity correlation coefficients less than .30 are not practically useful; correlations between .30 and .49 are practically significant; and correlations .50 or higher are very useful, but not common (Cashin, 1995).

positively and substantially correlated with effectiveness ratings of former students, effectiveness ratings of instructors' peers, and ratings of instructors by external observers. Student ratings of teaching effectiveness are not substantially correlated with administrators' ratings or instructors' self-ratings (Feldman, 1989b).

Expressiveness of Instructors

Critical assessment of the empirical evidence indicating that student ratings of instruction validly measure teaching effectiveness requires examination of a number of potentially biasing factors that may confound the ratings-achievement relationship. A prime candidate is the expressiveness of the instructor, which burst on the scene in 1973 with the publication in the *Journal of Medical Education* of a study authored by faculty members in the Schools of Medicine at the University of Southern California and Southern Illinois University.

The study (Naftulin, Ware, & Donnelly, 1973) investigated this hypothesis: "Given a sufficiently impressive lecturer, persons participating in a new learning situation can feel satisfied that they have learned when irrelevant, conflicting, and meaningless content is conveyed by the lecturer." Eleven psychiatrists, psychologists, and social work educators were the audience for the lecture. A professional actor was coached to present a lecture and conduct a question and answer session with an excessive use of doubletalk, neologisms, non sequiturs, and contradictory statements to be interspersed with humor and meaningless references to unrelated topics. A fictitious but impressive curriculum vitae was prepared. The actor was introduced to the group as the ambitious Dr. Myron L. Fox, an authority on the application of mathematics to human behavior. Dr. Fox presented a one-hour lecture with content purposely designed to be irrelevant, conflicting, and meaningless followed by a half-hour discussion period that was, in the authors' words, "hardly more substantive."

Following the presentation and discussion period, audience members completed an evaluation instrument consisting of seven questions to be answered "yes" or "no," indicating their satisfaction with the lecture and discussion. The questions pertained to the lecturer's interest in the subject, the use of examples, organization of content, stimulation of thinking, and whether material was presented in an interesting way. The lecture and discussion were video taped and, using the video tape, the authors repeated the study with another group of mental health educators followed by a group of educators and administrators enrolled in a graduate educational philosophy course.

Participants' responses to the questions were overwhelmingly "yes," indicating a high degree of satisfaction. One person even reported having read the lecturer's fictitious publications. The authors concluded that students can be "seduced into feeling satisfied that they have learned" when no substantive content is presented charismatically and "student satisfaction with learning may represent little more than the illusion of having learned." Thus two terms entered the lexicon of the evaluation of teaching – educational seduction and the Dr. Fox Effect – as a threat to the validity of student ratings as an index of teaching effectiveness.

Similar to the California professors' study claiming a negative relationship between student ratings and achievement, the Dr. Fox study was championed by those who questioned the

validity of student ratings. Methodological flaws of the study were noted – the most obvious being the absence of a control group, prompting the question “Irrelevant, conflicting, and meaningless content delivered charismatically compared to what?” Critiques of the study also pointed out that none of the post-lecture questions measured the knowledge acquired by lecture participants or sought their perceptions about their learning gain (Kaplan, 1974).

Correcting the methodological deficiencies of the original Dr. Fox study, researchers at Southern Illinois University designed an internally valid research paradigm to assess the influence of an instructor’s expressiveness – now called educational seduction – on student ratings. The new design, which became known as the “Dr. Fox paradigm,” examines the main and interaction effects of seductiveness and content covered on students’ learning and their ratings of teaching effectiveness. The design manipulates two levels of expressiveness (seduction) – high or low -- reflecting differences in vocal inflection, friendliness, charisma, humor, and personality. Also manipulated are three levels of content coverage described by the number of substantive teaching points covered – high coverage, medium coverage, and low coverage. Achievement is measured by tests; instructional effectiveness is quantified by students’ responses to scale items pertaining to a variety of lecturer behaviors and student outcomes.

Analyses of the studies conducted by the Southern Illinois University researchers (Ware & Williams, 1975; Williams & Ware, 1977) and by other researchers (Abrami, Leventhal, & Perry, 1982) have resulted in the following two conclusions regarding the effects of educational seduction (expressiveness) and content coverage on achievement and student ratings of teaching effectiveness:

- Expressiveness manipulations have substantial impact on overall students’ evaluations of teaching and small effects on achievement.
- Content manipulations have substantial effects on achievement and small effects on students’ evaluations.

The issue of the validity of student ratings as evidence informing tenure and promotion decisions reappeared in 1997 with the publication of a study conducted by two professors at Cornell University (Williams & Ceci, 1997). One of the authors, who regularly taught an undergraduate course in developmental psychology during both the fall and spring semesters, participated in a teaching skills workshop during the inter-session. An “enthusiastic” presentation style was taught during the workshop – speak with more pitch variability and use more hand gestures. The professors designed a study to investigate whether student ratings of the professor’s teaching during the spring semester could be improved by changing the “stylistic aspects of presentation” with no change in the content of the course.

At the end of the spring semester, students anonymously rated the instructor and the course using the same 10-item instrument that had been completed by students during the fall semester. The students’ overall effectiveness rating for the instructor for the fall semester was “average;” for the spring semester the instructor’s overall effectiveness rating was “good.” Students overall ratings for the course showed a similar increase. Persons who question the validity of student ratings of instruction especially highlight students’ responses to one item on the rating instrument: “How would you rate the text for this course?” The same text was used

both semesters. Fall semester students rated the text “poor;” spring semester students rated the text “average.” The quiz and test points earned and students’ final grades in the course did not differ for the two semesters. The Cornell professors emphasized what is most meaningful about the results is “the *magnitude* of the changes in students’ evaluations due to a content-free stylistic change by the instructor, and the *challenge* this poses to widespread assumptions about the validity of student ratings.”

Grading Leniency

Another potential threat to the interpretation of student ratings as an indicator of teaching effectiveness is the grading practice of instructors. The issue is whether an instructor’s grading leniency or strictness influences students’ evaluations (Greenwald, 1997).

Research has established that students’ evaluations correlate positively with course grades (Greenwald & Gillmore, 1997a). The best estimate of the magnitude of the relationship is about .20 (Marsh & Dunkin, 1997). A grading-leniency hypothesis proposes that instructors who give higher than expected grades will be rewarded with higher than deserved student ratings thereby resulting in a serious bias in interpreting student ratings as a measure of teaching effectiveness. A counter hypothesis – the validity hypothesis – contends that higher grades reflect greater learning by students, hence a positive correlation between students’ achievement and their evaluation of instruction supports the validity of student ratings. Then there is the students’ characteristics hypothesis stipulating that preexisting student variables such as motivation and prior interest may affect students’ achievement, their grades, and ratings of teaching effectiveness. Reviewers of the research have concluded that evidence supports the validity and students’ characteristics hypotheses with the qualification that the grading-leniency effect may produce some bias, which is not likely to be substantial (Marsh & Roche, 1997). Research at the University of Washington, however, indicates that grading leniency does influence student ratings and that statistical adjustment of students’ evaluative ratings to remove grading-leniency bias is called for (Greenwald & Gillmore, 1997a, 1997b).

Relationships Between Student Ratings and Other Factors

The interpretation of student ratings requires knowledge of the relationships between student ratings and characteristics of students, courses, and the circumstances under which student ratings are obtained. Empirical evidence documents the following relationships (Marsh & Roche, 1997; Cashin, 1995).

- Students who have higher prior interest in the subject generally rate instruction more favorably.
- Elective courses tend to be rated higher than required courses.
- Smaller enrollment classes are rated somewhat more favorably.
- Graduate level courses are rated somewhat more favorably than undergraduate courses; upper division undergraduate courses are rated higher than lower division courses.
- The rank of the instructor has little or no effect on student ratings.
- The sex of the instructor or student has little or no effect on student ratings.
- There is a weak tendency for students to rate courses in the arts and humanities higher than they rate social science, science, and mathematics courses.

- Somewhat higher ratings are obtained if students know that the ratings are to be used for tenure and promotion decisions.
- Somewhat higher ratings result when ratings are not anonymous and the instructor is present when students complete the ratings.
- Age and experience of the instructor are not correlated with student ratings.

The Teaching Effectiveness–Research Productivity Relationship

The proposition has been advanced that effective teaching and research productivity are closely allied; hence a faculty member’s research activities and publication record are valid indicators of teaching effectiveness. That proposition is not supported by empirical evidence, at least when teaching effectiveness is assessed by student ratings and research productivity is measured by publication records.

During the 1960s and early 1970s, two research articles were published reporting no significant relationship, statistical or practical, between extensiveness of publishing and students’ evaluations of teaching for faculty members at the University of Washington (Voeks, 1962) and Carnegie-Mellon University (Hayes, 1971). In the early 1980s an evaluation specialist at Educational Testing Service reported an analysis of the relationship between students’ overall effectiveness ratings and research productivity of faculty members in 70 colleges and universities. It was concluded that the teaching effectiveness-research productivity relationship “is either nonexistent or too modest to conclude that one necessarily enhances the other” (Centra, 1983). A meta-analysis of 29 studies of the teaching effectiveness-research productivity relationship reported in 1987 (Feldman, 1987) indicated an average correlation of $r = .12$. A 1996 meta-analysis of 58 studies reported an average correlation of $r = .06$ between the number of publications and teaching effectiveness ratings (Hattie & Marsh, 1996). The authors of both studies concluded that for all practical purposes, research productivity and teaching effectiveness are “essentially unrelated.”

Summary

At the beginning of the paper, four questions were posed for organizing the research about the use of student ratings in personnel decision making in higher education. I present the following summary statements.

- Students’ evaluations of teaching are reliable and stable; primarily the function of the instructor who teaches a course rather than the course that is taught (Marsh & Roche, 1997).
- Student ratings of instruction are relatively valid indicators of teaching effectiveness; however administrative, instructor, and course characteristics can influence the ratings (d’Apollonia & Abrami, 1997). University of Michigan Professor of Psychology Wilbert J. McKeachie, an internationally recognized scholar on college teaching, wrote the capstone article in a five-article series titled “Student Ratings of Professors’ in the November 1997 issue of the *American Psychologist*. The articles, written by authoritative experts on student ratings in American, Canadian, and Australian universities, report, critique, and synthesize the voluminous research since the early 1970s regarding the validity, interpretation, and use of student ratings (Greenwald, 1997). Professor

McKeachie indicated that he agrees with the other authors that student ratings “are the single most valid source of data on teaching effectiveness” (McKeachie, 1997). He also made three additional very important points. First, student ratings are not perfectly correlated with student learning; second, student ratings should be supplemented with other evidence of teaching effectiveness; and third, the major validity problem is in the use of students’ evaluations of instruction by colleagues, promotion and tenure committees, and administrators when making decisions about tenure, promotion, or merit pay. In addition to student ratings data, it is essential that administrators and promotion and tenure committees review evidence about the appropriateness, accuracy, currency, and rigor of course content – best judged by peers, not students – when making personnel decisions.

- Of the potential biasing variables for interpreting student ratings, instructors’ grading practices may be the most troublesome (Greenwald & Gillmore, 1997). Remember also what has been learned from the internally valid research about the “Dr. Fox Effect.” Content covered is more important in influencing achievement than is expressiveness; an instructor’s expressiveness has more influence on students’ satisfaction with instruction than it does on their achievement.
- And lastly, the evidence clearly indicates that research productivity quantified by publication records provides neither favorable nor unfavorable evidence about teaching effectiveness.

I conclude with this postscript, drawn from the Cornell professor’s claim that his changing to a more enthusiastic teaching style – use of hand gestures and voice pitch – resulted in students rating the quality of the text “average” in the spring semester when comparable students enrolled in the course during the fall semester rated the same text “poor.” If an enthusiastic presentation style can single handedly produce higher student ratings for the quality of a text book, perhaps I should have sought coaching to use hand gestures and vary voice pitch, thereby generating higher satisfaction ratings by members of Torch for the dinner we will enjoy soon – say a rating of “excellent” rather than “good.” Whether Torch members’ ratings of the quality of dinner validly describe the nutritional value of the food, I do not know. I am certain, however, that effective teaching that enhances students’ learning is more complex than voice pitch and hand gestures.

References

- Abrami P. C., Leventhal, L., & Perry, R. P. (1982). Educational seduction. *Review of Educational Research*, 52(3), 446-464.
- Cashin, W. E. (1995). *Student ratings of teaching: The research revisited*. (Paper No. 32). Kansas State University, Center for Faculty Evaluation & Development.
- Centra, J. A. (1983). Research productivity and teaching effectiveness. *Research in Higher Education*, 18(4), 379-389.
- Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, 51(3), 281-309.

- Cohen, P. A. (1990, Fall). Bringing research into practice. In M. Theall & J. Franklin (Eds.). *Student Ratings of Instruction* (pp. 123-132). New Directions for Teaching and Learning, No 43. San Francisco: Jossey-Bass.
- d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist*, 52(11), 1198-1208.
- 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(3), 227-298.
- Feldman, K. A. (1989a). 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(6), 583-645.
- Feldman, K. A. (1989b). 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(2), 137-194.
- Frey, P. W. (1973, October 5). Student ratings of teaching: Validity of several rating factors. *Science*, 182, 83-85.
- Gessner, P. K. (1973, May 11). Evaluation of instruction. *Science*, 180, 566-570.
- Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, 52(11), 1182-1186.
- Greenwald, A. G., & Gillmore, G. M. (1997a). Grading leniency is a removable contaminant of student ratings. *American Psychologist*, 52(11), 1209-1217.
- Greenwald, A. G., & Gillmore, G. M. (1997b). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology*, 89(4), 743-751.
- Haskell, R. E. (1997). Academic freedom, tenure, and student evaluation of faculty: Galloping polls in the 21st century. *Educational Policy Analysis Archives*, 5(6). Retrieved from <http://epaa.asu.edu/epaa/v5n6.html>
- Hattie, J., & Marsh, H. W. (1996). The relationship between research and teaching: A meta-analysis. *Review of Educational Research*, 66(4), 507-542.
- Hayes, J. R. (1971, April 16). Research, teaching, and faculty fate. *Science*, 172, 227-230.
- Kaplan, R. M. (1974). Reflections on the Doctor Fox paradigm. *Journal of Medical Education*, 49(3), 310-312.

- Marchase, T. (1997, September/October). Student evaluations of teaching. *Change*, 29(5), 4.
- Marsh, H. W., & Dunkin, M. J. (1997). Students' evaluation of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.). *Effective Teaching in Higher Education: Research and Practice* (pp. 241-320). New York: Agathon Press.
- Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist*, 52(11), 1187-1197.
- McKeachie, W. J. (1997). Student ratings. *American Psychologist*, 52(11) 1218-1225.
- Nuftulin, D. H., Ware, J. E., & Donnelly, F. A. (1973). The Doctor Fox lecture: A paradigm of educational seduction. *Journal of Medical Education*, 48(7), 630-635.
- Rodin, M., & Rodin, B. (1972, September 29). Student evaluations of teachers. *Science*, 177, 1164-1166.
- Seldin, P. (1999). Current practices – good and bad – nationally. In P. Seldin and Associates (Ed.). *Changing Practices in Evaluating Teaching* (pp. 1-24). Bolton, MA: Anker Publishing Co.
- Sheehan, D. S. (1975). On the invalidity of student ratings for administrative personnel decisions. *The Journal of Higher Education*, 46(6), 687-700.
- Trout, P. A. (1997, September/October). What the numbers mean. *Change*, 29(5), 25-30.
- Voeks, V. W. (1962). Publications and teaching effectiveness: A search for some relationship. *The Journal of Higher Education*, 33(4), 212-218.
- Ware, J. E., & Williams, R. G. (1975). The Dr. Fox effect: A study of lecturer effectiveness and ratings of instruction. *Journal of Medical Education*, 50(2), 149-156.
- Williams, R. G., & Ware, J. E. (1977). An extended visit with Dr. Fox: Validity of student satisfaction with instruction ratings after repeated exposures to a lecturer. *American Educational Research Journal*. 14(4), 449-457.
- Williams, W. M., & Ceci, S. J. (1977, September/October). How'm I doing? Problems with student ratings of instructors and courses. *Change*, 29(5), 13-23.
- Wilson, R. (1998, January 16). New research casts doubt on value of student evaluation of professors. *The Chronicle of Higher Education*, pp. A12-14.