Less-Than-Perfect Judges: Evaluating Student Evaluations

by Susanna Calkins and Marina Micari

Student ratings,[1] first introduced into the college classroom in the late 1920s, have long symbolized the often uneasy relationship between undergraduates and their professors.[2] Originally intended as an impartial and scientific means to gauge teaching performance, by the early 1960s, student ratings had become a site of anxiety and often bitter contest between faculty and students, and between faculty and administrators. Critical forces—such as emerging federal legislation that called for improved teaching in higher education, increasing demands for accountability in higher education, the increasing use of student ratings in personnel and curriculum decisions, the gradual democratization of the nation's campuses, and a developing consumerism in the nation's students—wrought tension and dissent within the higher education community. These tensions have manifested themselves in the debate over student ratings in different ways, often fixating on notions of validity, but revealing underlying complex challenges to the traditional structures of power and authority informing the changing faculty-undergraduate encounter.

Certainly, research about student ratings has been exhaustive over the last 50 years. Several thousand studies have been published, many concerned with issues of validity and reliability of evaluation instruments.[3] Yet, few studies have tackled the issue of student ratings in the context of the relationship between students and faculty and the place of higher education in society.[4] In this article, we examine the

Susanna Calkins *received her Ph.D. in European History from Purdue University and is an associate director at the Searle Center for Teaching Excellence at Northwestern University. She teaches in the Master of Science Higher Education Program and is co-author of* Learning and Teaching in Higher Education: The Reflective Professional. **Marina Micari** *is also an associate director at the Searle Center for Teaching Excellence. She has published in a variety of education journals on such topics as employee learning, undergraduate development, dynamics of small-group learning, and the impact of small-group learning on underrepresented students.*

academic discourse surrounding formal student ratings beginning when they first emerged, and focusing on the last fifty years.

We first trace their development from near-private to semi-public communications, and the resulting discomfort that developed among faculty. Then, drawing on a wide selection of scholarly journal articles and popular commentaries, we examine how the notion of validity served as a weapon for faculty who wished to discredit the legitimacy of the student voice outright, as well as those who wished to discredit the tools for conveying that student voice, or to protest the undue use of that student voice in personnel decisions. At the same time, we suggest validity was also used to protect students as well as faculty, and legitimate their opinions and concerns. We look, too, at the strategies faculty have used to retain their own power and voice within a shifting and often poorly marked academic terrain.

We first review the historical use of student ratings of teachers, tracing their development from near-private to semi-public communications.

In the sections that follow, we trace the history of and scholarly discourse surrounding the use of student evaluation of teaching in U.S. institutions of higher education. By emphasizing that discourse, we bring light to the ways in which the academic community has viewed student ratings over time, and more significantly, the anxieties surrounding and corresponding attacks on the practice of students evaluating teachers, as well as the more recent attempts to resolve these tensions.

The first formal student ratings systems were developed at Purdue University in the mid-1920s as part of a systematic inquiry into the traits associated with good teaching, which included "fairness in grading," "stimulating intellectual curiosity," and "personal peculiarities."[5] In the interwar period, Purdue, like other public research universities, was responding to the larger federal push toward scientific research that followed World War I. Social scientists were starting to apply scientific methods to disciplines outside of the sciences, and there was a growing belief that such methods could solve practical problems associated with college teaching.[6] Using the results from his Purdue Rating Scale for Instructors, Herman Remmers, a psychology professor, pioneered investigations in student ratings. Among other things, Remmers examined the degree to which students' judgments about a course agreed with those of their peers and administrators, the degree to which students changed their minds about teachers after they left college, and whether the judgments of poorer students should be considered.[7]

While student ratings systems were not immediately widespread, by the early 1940s, many institutions of higher education across the nation had begun to implement some form of student ratings. But there was little formal accountability at this time, and few questions arose about instructors' teaching or grading methods. Within the curriculum, faculty usually could select course content and construct exams as they deemed appropriate for the topic. The figure of the (usually white male) faculty member at the university was detached from his students: a scholar who expounded his authority, knowledge, and expertise from a distant podium.[8] The instructors' authority and control over their teaching was bolstered, too, by the American Association of American Professors' (AAUP) initial statements on academic freedom and tenure.[9] Moreover, with a shortage of qualified professors in the U.S., university officials shielded their teachers by protecting their academic rights, including their autonomy in the classroom. During the immediate postwar period, faculty members still could teach largely as they wished, letting students "sink or swim," even as enrollment numbers swelled with the passing of the G.I. Bill, and lecture halls overflowed.[10]

By the early 1940s, many institutions of higher education across the nation had begun to implement some form of student ratings.

Increasingly, however, as the faculty committee structure began to develop in the 1940s and 1950s, faculty members gained new opportunities to participate in institutional governance, the appointment and promotion process, and in forging educational policy.[11] In one sense, this new type of administrative positioning gave faculty greater power and voice, but in another sense, it pulled the faculty out of their semi-isolation. Administrators and faculty senates began to require student ratings systems, with the idea that the students' comments would play a role in salary, promotion and tenure decisions. The comments were largely a private one-way communication from students to the instructor, but selected faculty and administrators could gain access to the student comments in order to make personnel decisions. While not yet a mechanism of control, arguably student ratings were starting to curtail the faculty's autonomy in the classroom. But with little accountability to the university for upholding any teaching standards, professors could still dismiss student comments as they wished.

Faculty complaisance about academic freedom, and the professors' nearly untouchable role in institutions of higher education, was shaken to the core by the Cold War rhetoric and McCarthy proceedings of the 1950s. The government—always suspicious of colleges and universities as hotbeds of dissent and radicalism—began to target faculty with suspected Communist leanings. Hundreds of faculty careers ended abruptly, quietly, and without due process.[12] Faculty found their research and teaching activities under new scrutiny. Moreover, after the Soviet Union successfully launched *Sputnik* in 1957, the U.S. government, grimly determined to keep the U.S. competitive at a global level, passed the *National Defense of Education Act* (1958) to bolster education, especially in fields of science, technology, and foreign language.[13] Improving teaching became a matter of national concern.

As faculty power declined somewhat in the 1950s, student power grew and became more organized. By the end of the decade, too, the "impersonality of the multiversity"[14] was brewing discontent among students tired of being "just a number" in a cold bureaucracy. Students had little interaction with faculty, and less with the administration, and began to feel—and more importantly, question—this detachment and lack of voice. New democratizing influences, new perceptions of mass accessibility, and a new consumer ethos had begun to creep into American higher education.[15]

Much has been written about how college students, baby boomers, questioned authority and changed campus culture in the 1960s and 1970s. And as students in the 1960s began to publicly question the effectiveness of their instruction, especially large lecture classes, as well as to comment on the teaching abilities of their professors and graduate teaching assistants, student ratings, in many ways, symbolized their struggle for self-determination and control.[16] As contemporary observers noted, "One facet of recent campus unrest is the growing tendency of students to question and challenge traditional approaches to teaching."[17]

In the 1960s and 1970s, too, for the first time, student groups were often given control of the student rating or course evaluation process. This gave students an unprecedented point of power, particularly when they could publish and circulate

student ratings with little monitoring or oversight.[18] For example, psychologists Wendy Williams and Stephen Ceci recalled how student groups would administer and collect student ratings at their respective colleges, annotate them with often caustic or snide remarks, and circulate the annotated ratings widely through the university community, so that they passed easily through the hands of department chairs, deans, and faculty colleagues.[19]

In another instance, a group of student leaders rallied their fellow seniors to each write a paragraph about an ineffective professor with whom they had taken a class.[20] A student committee collated the write-ups. While it is not clear whether the professors were identified by name in the students' initial summary, an article detailing the ineffective practices, quoting descriptions and actions of unnamed professors, was published in a leading journal on college teaching soon after. Here, the student committee identified seven vices, which included instructor lack of preparation and organization, a lack of enthusiasm about lecture presentation, lack of effective teaching techniques, inability to communicate clearly, lack of knowledge about students, and lack of adequate subject knowledge.[21] While students may long have griped about such issues within the confines of the college walls, here they were making their concerns about teaching effectiveness more public.

In the 1960s and 1970s, for the first time, student groups were often given control of the student rating or course evaluation process.

Student ratings, a purportedly private and confidential communication from students to professor, were becoming semi-public. Not only were they a vehicle for students to communicate to one another about which teachers' courses to take, and which to avoid; students were also communicating the idea that their opinions about quality teaching were valid and meaningful. This view was confirmed by the AAUP's *Statement on Teaching Evaluation* (1974) which asserted that "Student perceptions are a prime source of information from those who must be affected if learning is to take place."[22]

For the first time, faculty were asked to explain poor or inconsistent ratings, and generally to defend their teaching practices and choice of subject material.[23] This gave rise to an anxious tenor in the conversation over student ratings. In effect, they had become a mallet used to chip away at the faculty's perceived authority and protected status within the ivory tower.

In negotiating this shifting academic terrain, faculty appear to have responded in several ways. Some, no doubt, opted to do nothing. This strategy could only work, of course, for faculty with nothing to lose: tenured faculty confident in their rank and status at the university. Faculty who generally received good ratings might also stay out of the fray; after all, the debate was over "bad" teachers, not them.[24] Others certainly opted to use the student ratings to improve (a strategy we discuss in more depth below, as it was not a very common tactic until recent decades). A portion may have simply sought to manipulate the system, by "buying" good student ratings with good grades, easy tests, "fun" teaching activities, or other tactics designed to curry good favor among their students.[25]

But many faculty opted for a different strategy altogether: go on attack. Shaken by the uneasy world around them, faculty challenged student ratings, primarily on the notion of validity, underscoring both their own uneasiness over the foothold students had found in higher education, and their own increasingly tenuous space at the university, particularly for untenured faculty.

Questions concerning the research validity of student ratings dominated the scholarly discourse on student evaluations throughout the 1970s.

In particular, questions concerning the research validity of student ratings dominated the scholarly discourse on student evaluations throughout the 1970s. Skeptics analyzed instrumentation, bias based on class size, class type, and expected grade, and lack of a relationship between teaching evaluation and student learning outcomes, citing all of these as threats to the validity of the student-evaluation research enterprise.[26] Study authors regularly used their results to argue for the ultimate illegitimacy of the practice of students rating their teachers.

The now-famous "Dr. Fox experiment," in which a charismatic actor giving a nonsense lecture was rated highly by a well-educated audience, reignited any smoldering doubts about the validity of evaluations.[27] Another widely cited study showed that students tend to assign high ratings to teachers from whom they learn the least, promoting debate about whether students could, in fact, be trusted to make sound judgments about the quality of the teaching they encounter. Rather than the student —"the primary consumer of the teaching product"— being in the best position to evaluate teaching, the researchers claimed that students "are less than perfect judges of teaching effectiveness if the latter is measured by how much they have learned."[28] The nature of these studies reflected the anxiety faculty experienced in opening their professional practice to critique from students, but these concerns may also have prompted some faculty and departments to limit the use of student ratings.[29]

Researchers continued to focus on the validity of course evaluations into the later part of the 1970s and 1980s, taking an increasingly sophisticated bent. In contrast to earlier studies looking at just a handful of factors (such as the correlation between evaluations and student performance[30]), studies from this later period began to take into consideration multiple factors—sometimes using experimental designs—and to examine the relationships among these, in considering the validity of the student-evaluation system. Researchers examined questions such as the impact of stringency of grading and test frequency on evaluations, the time at which evaluation is done, the influence of particular components of evaluation forms, and the impact of such factors as rater anonymity and the teacher's presence in or absence from the room,[31] among numerous other factors. This increasing complexity might be seen as part of the normal development of any area of social-science research, but might also have emerged out of frustration, as a response to what had become something of a dead-end research direction: There were many complaints about student ratings and proposed reasons to end them, but with little chance of this happening. Broadening the analyses to examine more complex relationships among factors in teaching-evaluation practice would at least open new windows and provide openings for new directions in the conversation.

There were many complaints about student ratings and proposed reasons to end them, but little chance of this happening.

In the 1980s, student ratings, like other long-standing traditions in higher education, came under a new critical lens, scrutinized for their underlying messages of power and authority. In particular, the question of whether and how gender—the instructor's and the students'—might play a role in student ratings surfaced more prominently. More female professors than ever before comprised the professoriate in the 1980s, often experiencing a "chilly climate" among their male peers and in their classrooms.[32] Research suggested that students held "contradictory and unrealistic expectations" of their female professors. Baker and Copp noted:

> Because women's culturally defined gender status clashes with their occupational status as professors, students may hold contradictory and unrealistic expectations of them. These contradictions may make it hard for women faculty members to receive outstanding teaching evaluations, because students judge women by their gender performance, first, and by their teaching second.[33]

Other contemporary studies found that students expect women to be warmer than men; that male students give female professors lower ratings than they gave male professors; that male, more than female, students are influenced by instructor gender; that students do not want men to wear pink shirts when they teach (implying that students are uncomfortable with male instructors who do not conform to traditional notions of masculinity); and even that male and female instructors have different understandings of what is "good" teaching.[34] Disclosure of these perceived student biases—however accurate—seemed to give critics further ammunition: Non-male, non-heterosexual instructors could hardly get a fair shake when being rated by their students, and as such, student ratings could have no place in personnel decisions. The gendered subtext underlying these challenges to the validity of student ratings continued to pull the student voice into question, again questioning the student's ability to accurately assess teaching.

Amid the more strident challenges to the validity of course evaluations, a quieter faculty voice had begun to express support for student evaluations.

Interestingly, race and ethnicity did not become a platform for questioning the validity of student ratings for nearly two more decades—the first published critiques did not appear until the turn of the millennium (a point we return to below). We can only surmise a partial explanation for this silence: Since the vast majority faculty in the 1970s and 1980s were white,[35] researchers may have considered sample sizes of underrepresented faculty on their campuses to be too small or even nonexistent.[36] We should also note that we could find no studies from this time period indicating that African American students, or students from any other underrepresented groups, rated their instructors differently from their white counterparts. Why researchers were not investigating the effects of race in student ratings we can not say, but clearly there was no critique launched on this particular front for twenty more years.

Amid the more strident challenges to the validity of course evaluations, a quieter faculty voice had begun to express support for student evaluations of teaching in the early 1970s, growing in strength through the 1980s and beyond. This voice asked: How can one learn from student ratings and use that knowledge to improve teaching? College teachers and those studying teaching had begun to realize that focusing solely on research validity of teaching evaluation was reducing the problem to meaninglessness. This voice whispered about utility and improvement, about seeking new strategies to circumvent the validity barrier, about constructively dealing with problems related to the instructor and instruction. This new way of thinking about student ratings represents on a small scale the far larger paradigm shift occurring in the latter part of the 20th century: the shift from an instructional paradigm (one that puts the teacher and instruction at the center of teaching) to a learning paradigm (one that places the learner at the center of teaching).[37]

Questions turned from whether and how evaluation tools may or may not be telling the truth to how evaluation can be conducted in more meaningful ways.[38] Much of this conversation focused on developing evaluation instruments and broader practices that would provide meaningful information—that is, information that can allow instructors to improve their teaching. Suggestions for improvement included developing instruments that would provide more meaningful data, for instance data reflecting aspects of teaching valued by faculty and administrators as well as students; using alternative constructs in instruments that more accurately reflect good teaching; rounding out teaching evaluation so that data would be triangulated, for instance by using observations in addition to surveys; and including instructors in the development of survey instruments.[39]

Questions turned from whether and how evaluation tools may or may not be telling the truth to how evaluation can be conducted in more meaningful ways.

Despite the positive direction of this strand of conversation, the commentary still revealed a heightened awareness of the "rival purposes" of student ratings, that is, the conflict between the course evaluation as a factor in tenure and promotion decisions on the one hand, and as a tool for promoting improvements in teaching on the other.[40] These dual purposes now identified and labeled, the conversation branched into two directions: one focusing on how to make fairer administrative decisions about faculty promotion and tenure, and the other exploring how to increase the likelihood that evaluations will help lead to teaching improvement— and improved student learning, by extension. In the former case, calls for including more systematic collection of data in student ratings—using standardized instruments and trained observers; more personal information, such as case histories or qualitative information from students; and teaching portfolios that complement data[41]—sent the message that student ratings could be effectively used for administrative decision-making purposes, provided care was taken to develop evaluation systems appropriate to this goal. Academic departments were urged to implement practices that would help faculty become better teachers through the evaluation process, for instance by using midterm feedback, working with faculty consultants, adding a self-evaluation component to evaluations, and using group-discussion techniques, in addition to to the use of traditional surveys to collect student feedback.[42]

While there was no immediate reconciliation of these rival purposes, an expanded focus on the people doing the improving—the instructors and students themselves—emerged in the 1990s and 2000s, perhaps prompted by the new emphasis on improving teaching and learning. Investigations of issues like faculty members' psychological traits and how these might be related to student ratings, faculty attitudes about evaluation, students' beliefs about the evaluation system, students' perceptions of their own role in the evaluation process, the impact of perceived instructor caring on learning and student ratings all entered the conversation, bringing a humanistic tenor to the discourse. As the conversation became more person-oriented, it increasingly took on concerns of student learning, with assessment of teaching now directly linked to assessment of learning.[43]

The conversation increasingly took on concerns of student learning, with assessment of teaching now directly linked to assessment of learning.

The emergence of online student ratings systems has ushered in new faculty concerns. Proponents argue that online ratings make the evaluation process less cumbersome, and that the swiftness and ease of viewing the comments would be worth the change.[44] On the other hand, since students can complete online evaluations on their time, and in their own space, challenges to the new process have focused on the implications of lower response rates. Critics have argued that (1) only those students with the proverbial axe to grind will fill out the ratings, (2) students will write only negative remarks; and (3) negative remarks will be greater in length than any positive remarks. Even though preliminary studies have suggested these beliefs to be false,[45] the underlying concern remains, and may stem from a loss of control over the system. Faculty who believe they can manipulate the process, for example, by refraining from returning major assignments until after the student ratings have been completed, or simply "buying" the students with overt forms of good will (e.g. passing out chocolate or pizza with the evaluation form), may perceive they now have less influence over the process.[46]

Where student ratings are viewable online to all members of the university community, faculty have expressed other concerns. Aware that their comments are going to be posted publicly, many students will simply address their remarks about the instruction and instructor to one another, at times encouraging their peers to take the class, but just as likely warning future students to take the course with caution, or even to stay away completely. A dialogue of sorts ensues, not between students and faculty necessarily, but between current and future students. Faculty are, in effect, being excluded from the discussion, and pushed, to some degree, into the position of bystander.

A more interesting development, however, is in how students have taken student ratings out of the hands of the faculty, and even the administration entirely. The introduction of new websites, such as ratemyprofessor.com, which allow students to comment publicly and openly on the Internet about their teachers, have enjoyed great success.[47] These sites allow anyone to comment on anyone's teaching; conceivably even people who have never even met the instructor can comment favorably or unfavorably on their teaching prowess. But although the chaotic nature of unauthorized student ratings may cause distress to faculty,[48] there is little evidence to date that universities even look at such sites.

Clearly, the specter of what those little online chili peppers—the rating schema preferred by one site to indicate professorial hotness—can do may add nuance to faculty concerns about student evaluations, but the overall tenor of the conversation seems to be shifting once again. The chili pepper, often mocked in informal commentary, seems for faculty to be a symbol of the absurdity of these student comments. These student voices are not sanctioned by the university, and thus for now, lack legitimacy. The current academic sense is that, while colorful and expressive, they do not constitute a "real" voice, and are not yet a "real" threat. Though such student voices are still operating on the margins, faculty seem to have some sense—perhaps based on history—that those margins can become a new center.

Although the chaotic nature of unauthorized student ratings may cause distress to faculty, there is little evidence to date that universities even look at such sites.

Student ratings, clearly, have long been a site for competing tensions between faculty and students, and faculty and administrators. At a time when instructors were not always held accountable for their teaching, student comments could be ridiculed, dismissed, or hidden from view.[49] Over time, as the process became increasingly public, visible, and more critical in personnel decisions, the dynamics of the evaluation of the faculty-student relationship have appeared to shift.

The contention that student evaluations of teaching lack validity has no doubt provoked the most heated argument on the topic of student evaluations, both in the literature and in faculty offices across the country. This concern would be at least somewhat muted if student teaching evaluations were not considered in tenure-and-promotion decisions. Certainly, there are legitimate concerns about general bias in student evaluations of teaching. Charismatic teachers, for instance, are typically rated more highly on overall ability, and studies on gender point to clear tendencies for male and female instructors to be judged differently for the same behaviors. In this case, there appear to be intervening factors, such as course type, discipline, and instructor personality, but nevertheless gender certainly can play a role in how a teacher is evaluated. Ethnicity and language ability have also been found to correlate with students' evaluations of teachers,[50] as has sexual orientation.[51] Attractiveness, too, has been found to be related to student evaluations of teachers, with an even greater impact for male than for female teachers.[52]

Interestingly, there has been less discussion in the literature of bias on the basis of professor ethnicity or race than on the basis of gender, and most of the studies published on ethnicity and teaching evaluations did not appear until the 2000s. These studies have found mixed results,[53] and it appears that factors such as course type and student ethnicity may confound the findings.[54] But clearly there are some patterns in students' evaluations of teachers based on ethnicity.

The potential for bias, however, in no way renders the evaluations useless. The answer is not to do away with them, but rather to use them wisely. This means, in part, that faculty, tenure committees, and university administrators should be made aware of the potential for bias and how it might influence evaluations, and faculty should be supported in presenting their evaluations in light of potential bias. They might take a lesson, for example, from institutions that provide a standard note on teaching evaluation results explaining that bias can affect the results.[55]

It is the design of teaching evaluations that should be central to the debate, not the inherent validity of the enterprise itself.

In a sense, however, the longstanding debate over teaching evaluation validity misses the point. Teaching evaluations are as valid as their designs allow them to be. Many are not really measuring teaching. It is the design of teaching evaluations that should be central to the debate, not the inherent validity of the enterprise itself. Evaluation forms that ask little more than whether the student liked the course and thought highly of the instructor do run the risk of measuring personality and charisma more than the ability to promote learning. But when students are asked questions that probe learning (such as whether their attitudes or beliefs were changed, whether they understand connections more fully, whether they feel more confident in their ability to tackle problems of the field) and the teacher's approach to facilitating learning (such as whether the teacher answered students' questions, invited students to office hours, and promoted student engagement), the evaluation comes much closer to measuring teaching skill.[56]

The opening and seeming democratization of the teaching evaluation process has provoked significant tension and debate within higher education: Is it a positive change, liberating and empowering students and promoting higher-quality teaching, or does it only further render the student a consumer and the teacher, and university by extension, a provider of products to be tailored to the buyer's desires? Intention and design determine this answer.

Ultimately, we hope we have complicated the understanding of the critical role that student ratings have played in the faculty-student encounter in higher education: They are simultaneously a source of ongoing tensions and conflicts at the university, a symbol of shifting power relationships, a medium of exchange, a representation of diverse student voices, and finally, a means of communication that should help to keep this ongoing conversation alive well into the future.

ENDNOTES

1. We found that terms such as "teaching evaluations," "course evaluations," "student evaluations," "student ratings," and "teacher ratings" are used somewhat interchangeably throughout the literature; we use the general term "student ratings" to broadly indicate standardized rating forms that ask students to anonymously comment on or rate various course factors. Traditionally, these are administered towards the end of a course.
2. Herman H. Remmers, *Learning, Effort, and Attitudes as Affected by Three Methods of Instruction in Elementary Psychology*, Monograph No. 21, (West Lafayette, IN: Purdue University Studies in Higher Education, 1933);

   Wilbert J. McKeachie, "Research on College Teaching: The Historical Background," *Journal of Educational Psychology* 82, no. 2 (1990): 189-200.
3. See for example A.C. Maney, "The Authoritarianism Dimension in Student Evaluations of Faculty," *Journal of Educational Sociology* 32, no. 5 (1959): 226-231; D. Solomon "Teacher Behavior Dimensions, Course Characteristics, and Student Evaluations of Teachers," *American Educational Research Journal* 3, no. 1 (1966): 35–47; M. Rodin, and B. Rodin, "Student Evaluations of Teachers," *Science*, 177, (1972): 1164–1166; P.A. Cohen, "Student Ratings of Instruction and Student Achievement: A meta-analysis of Multisection Validity," *Review of Educational Research* 51 (1981): 281–309; Herbert W. Marsh and L.A. Roche, "Making Students' Evaluations of Teaching Effectiveness Effective," *American Psychologist* 52 (1997): 1187–1197; M. Shevlin, P. Banyard, M. Davies, and M. Griffiths. "The Validity of Student Evaluation of Teaching in Higher Education: Love Me, Love My Lectures?" *Assessment and Evaluation in Higher Education* 25, no. 4 (2000): 397–405.
4. Some exceptions include Wendy M. Williams and Stephen J. Ceci, "'How'm I doing?' Problems with Student Ratings of Instructors and Courses," *Change* (1997): 13-23; Paul A. Trout, "What the Numbers Mean: Providing a Context for Numerical Student Evaluations of Courses," *Change* (1997): 24-30; Bob Algozzini, John Beattie, Marty Bray, Claudia Flowers, John Gretes, Lisa Howley, Ganesh Mohanty, and Fred Spooner, "Student Evaluation of College Teaching: A Practice in Search of Principles," *College Teaching* 52 (2004): 134-141.
5. Other traits were interest in subject, sympathetic attitude towards students, liberal and progressive attitude, presentation of subject matter, sense of proportion and humor, self-reliance and confidence, and personal appearance. J.D. Heilman and W.D. Armentrout, "The Rating of College Teachers on Ten Traits by Their Students," *Journal of Educational Psychology* 27 (1936): 197-216.
6. McKeachie, "Research," 189-191.
7. Ibid. 194-195.
8. J.R. Thelin, *A History of American Higher Education* (Baltimore: Johns Hopkins, 2004): 287.
9. American Association of American Professors (AAUP). (1925; 1940). "Statement of Principles on Academic Freedom and Tenure."
10. M. Trow, "From Mass Education to Universal Access: The American Advantage." In Altbach, *In Defense of American Higher Education*, (Baltimore: Johns Hopkins, 2001); Thelin, *History*, 262-268.
11. M.J. Finkelstein, *The American Academic Profession: A Synthesis of Social Scientific Inquiry Since*

World War II (Columbus: Ohio State University Press, 1984).

12. Thelin, *History*, 274-277.

13. National Defense of Education Act (1958).

14. Thelin, *History*, 307.

15. In an early study, for example, Joseph E. Grush and Frank Ostin, concluded that "college students are objective consumers of the teaching process and their judgments should be solicited to identify variables important for teacher effectiveness." In "The Student as Consumer of the Teaching Process," *American Educational Research Journal*, 12 (1975): 55.

16. Kenneth E. Eble, and Wilbert J. McKeachie, *Improving Undergraduate Education through Faculty Development*, (San Francisco: Jossey-Bass, 1985).

17. Elaine R. Parent, C. Edwin Vaughan and Keith Wharton, "A New Approach to Course Evaluation," *Journal of Higher Education* 42, no. 2 (1971): 133.

18. Williams and Ceci, "Problems," 13.

19. Ibid. 13.

20. The college in this article was not identified, and the author was identified only as a "nationally known scholar" writing under a pseudonym. Hampton, Everett, "Seven Teaching Sins as Seen by Seniors," *Improving College and University Teaching* 19, no 3 (1971): 248-249.

21. Everett, "Seven," 248-249.

22. American Association of University Professors Committee C on Teaching, Research and Publication, "Statement on Teaching Evaluation," *AAUP Bulletin* 60, no. 2 (1974): 166-170.

23. Williams and Ceci, "Problems," 13.

24. Although this is, of course, subject to debate. As Williams and Ceci point out, "A professor of organic chemistry, for example, might be forced to explain his lower-than-average ratings by pointing to the rigor and monotony inherent in learning organic chemistry, while a modern film teacher might bask in uniformly excellent ratings after showing five feature films over the semester." In "Problems," 13.

25. L. Aleamoni, "Student Rating Myths Versus Research Facts," *Journal of Personnel Evaluation in Education* 1, (1987): 111-119.

26. Donald A. Bligh, What's the Use of Lecturing? (Devon, England: Teaching Services Centre, University of Exeter, 1971); Gordon E. Greenwood, Charles M. Bridges, Jr., William B. Ware, James E. McLean, "Student Evaluation of College Teaching Behaviors Instrument: A Factor Analysis," *Journal of Higher Education* 44, no. 8 (1973): 596-604; Kathleen S. Crittenden, James L. Norr, Robert K. LeBailly, "Size of University Classes and Student Evaluation of Teaching," *Journal of Higher Education* 46, no. 4 (1975): 461-470; Kenneth Wood, Arnold S. Linsky, Murray A. Straus, "Class Size and Student Evaluations of Faculty," *Journal of Higher Education* 45, no. 7 (1974): 524-534; M. L. Silberman and J. S. Allender, "The Course Description: A Semiprojective Technique for Assessing Students' Reactions to College Classes," *Journal of Higher Education*,. 45, No. 6, (Jun., 1974): 450-457; C.R. Snyder and M. Clair, "Effects of Expected and Obtained Grades on Teacher Evaluation and Attribution of Performance," *Journal of Educational Psychology*, 68 (1976): 75-82.

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

28. Rodin and Rodin, "Student," 1164–1166.

29. J. W. Gustad, "Evaluation of Teaching Performance: Issues and Possibilities," In C. B. T. Lee (Ed.), Improving College Teaching (Washington, D.C.: American Council on Education, 1967); Costin, E, Greenough, W.T., and Menges, R.J., "Student Ratings of College Teaching: Reliability, Validity, And Usefulness," *Review of Educational Research* 41 (1971): 511-535.

30. Peter W. Frey, Dale W. Leonard and William W. Beatty, "Student Ratings of Instruction: Validation Research," *American Educational Research Journal*, 12 (4) (1975): 435-447; Sullivan,

A. M., and Skanes, G. R. Validity of student evaluation of teaching and the characteristics of successful instructors. *Journal of Educational Psychology* 66, (1974): 584-590.

31. R. W. Powell, "Grades, Learning, and Student Evaluation of Instruction," *Research in Higher Education* 7 (1977): 193-205.; Kenneth A. Feldman, "The Significance of Circumstances for College Students' Ratings of Their Teachers and Courses," *Research in Higher Education* 10, no. 2 (1979): 149-172.

32. Roberta. M. Hall and Bernice R. Sandler, The Classroom Climate: A Chilly One for Women? *Report of the Project on the Status and Education of Women*, (Washington, DC: Association of American Colleges, 1982).

33. Baker and Copp, "Gender Matters," 29.

34. See for example, S.A. Basow and N.T. Silberg, "Student Evaluations of College Professors: Are Female and Male Professors Rated Differently?" *Journal of Educational Psychology* 79 no. 3 (1987): 308–314; E. Kaschak, "Sex Bias in Student Evaluations of College Professors," *Psychology of Women Quarterly* 2 (1978): 235-243; Laura D. Goodwin and Ellen A. Stevens, "The Influence of Gender on University Faculty Members' Perceptions of 'Good' Teaching," *Journal of Higher Education* 64, no. 2, (1993): 166-185.

35. Robert J. Menges and William H. Exum, "Barriers to the Progress of Women and Minority Faculty" *The Journal of Higher Education*, 54, no. 2 (Mar.-Apr., 1983), pp. 123-144.

36. Even by 2000, JoAnn Miller and Marilyn Chamberlin noted in their study of gender and student ratings, they could not address race: "The sociology department where this study was conducted has only one Black man and one Black woman on the faculty. The extremely small representation of Blacks on the faculty precludes the inclusion of race in the research design" p. 286. In "Women Are Teachers, Men Are Professors: A Study of Student Perceptions," *Teaching Sociology*, 28, no. 4 (Oct., 2000), pp. 283-298.

37. R.B. Barr and J. Tagg, "From Teaching to Learning: A New Paradigm for Undergraduate Education," *Change*, 27 (1995): 12-25.

38. For example, see Parent, et al, "New Approach," 133-138; H. Richard Smock and Dale C. Brandenburg "A plan for the comprehensive evaluation of college teaching" *Journal of Higher Education* 49 no. 5, (1978): 489-503.

39. Thomas R. Wotruba and Penny L. Wright, "How to Develop a Teacher-Rating Instrument: A Research Approach," *Journal of Higher Education* 46, No. 6, (Nov. - Dec., 1975): 653-663; Gordon E. Greenwood and Howard J. Ramagli, Jr., "Alternatives to Student Ratings of College Teaching Source," *Journal of Higher Education* 51, no. 6, (1980): 673-684.

40. J.O. Derry, "Can Students' Ratings of Instruction Serve Rival Purposes?" *Journal of Higher Education*, 50 (1) (1979): 79-87; James O'Hanlon and Lynn Mortensen, "Making Teacher Evaluation Work," *Journal of Higher Education* 51, no. 6, (1980): 664-672.

41. Gaylord L. Thorne, Student Ratings of Instructors: From Scores to Administrative Decisions. *Journal of Higher Education* 51, no. 2, (1980): 207-214; Yi-Guang Lin, Wilbert J. McKeachie, David G. Tucker, "The Use of Student Ratings in Promotion Decisions," *Journal of Higher Education* 55, no. 5, (1984): 583-589; John A. Centra, "The Use of the Teaching Portfolio and Student Evaluations for Summative Evaluation," *Journal of Higher Education* 65, no. 5, (1994): 555-570.

42. J. U. Overall and Herbert W. Marsh, "Midterm Feedback from Students: Its Relationship to Instructional Improvement and Students' Cognitive and Affective Outcomes," *Journal of Educational Psychology* 71, no. 6 (1970): 856-865; Robert C. Wilson, "Improving Faculty Teaching: Effective Use of Student Evaluations and Consultants," *Journal of Higher Education* 57, no. 2 (1986): 196-211.

43. Christiane Brems, Michael R. Baldwin, Lisa Davis, Lorraine Namyniuk, "The Imposter Syndrome as Related to Teaching Evaluations and Advising Relationships of University Faculty Members," *Journal of Higher Education* 65, no. 2, (1994): 183-193; P.M. Simpson and J.A. Siguaw, "Student Evaluations of Teaching: an Exploratory Study of the Faculty Response,"

Journal of Marketing Education 22, no. 3 (2000): 199–213; F. Nasser and B. Fresko "Faculty Views of Student Evaluation of College Teaching," *Assessment and Evaluation in Higher Education*, 27, (2002): 187–198; James W. Marlin, Jr., "Student Perceptions of End-of-Course Evaluations," *Journal of Higher Education* 58, no. 6, (1987): 704-716; J. Sojka, A. K.Gupta and D. R. Deeter-Schmelz, "Student and faculty perceptions of student evaluations of teaching," *College Teaching* 50, no. 2 (2002): 44–49; William Cerbin, "The Course Portfolio as a Tool for Continuous Improvement of Teaching and Learning," *Journal on Excellence in College Teaching* 5, no. 1 (1994): 95-105.

44. Nedra Hardy, "Online Ratings: Fact and Fiction," *New Directions for Teaching and Learning*, 96 (2003): 32.
45. Hardy, "Online Ratings," 31-38.
46. See, for example, Paul Trout's accusation of faculty members throwing pizza parties before administering evaluations in "Flunking the Test: The Dismal Record of Student Evaluations," in *Academe* 86: 58-61, and Eugene Arden's measured response, "Should Students Evaluate Faculty Members?" *College Teaching* 50 (2002): 158-159.
47. Rebecca Herzig, "So Much Depends Upon a Red Chili Pepper," *Liberal Education* (2007): 26-31.
48. Herzig, "Chili Pepper," 26.
49. Stanley Fish, for example, has openly admitted to throwing evaluations in the trashcan rather than administering them to his students in "Who's in Charge Here?" *Chronicle of Higher Education* 51 (2005): C2-C3.
50. Thomas Galguera. "Students' Attitudes towards Teachers' Ethnicity, Bilinguality, and Gender." *Hispanic Journal of Behavioral Sciences* 20. 4, (1998): 411–429. D.S. Hamermesh & A. Parker. "Beauty in the Classroom: Instructors' pulchritude and putative pedagogical productivity," *Economics of Education Review* 24 (2005): 369-376.
51. Vanessa Ewing, Arthur Stukas, Jr., and Eugene Sheehan. "Student Prejudice against Gay Male and Lesbian Lecturers." *The Journal of Social Psychology* 143.5 (2003): 569–580.
52. Hammermesh & Parker, "Beauty in the Classroom,"369-376.
53. Jack Glascock and Thomas Ruggiero found only a negligible effect for professor ethnicity on student ratings, from among 486 students at a predominantly Hispanic institution (Glascock & Ruggiero, "The Relationship of Ethnicity and Sex to Professor Credibility at a Culturally Diverse University," *Communication Education* 55.2 [2006]: 197-207). Hammermesh and Parker did find a difference, however, with minority faculty rated less highly than majority, in 463 courses at the University of Texas in "Beauty in the Classroom," 369-376. On the other hand, Arnold Ho, Lotte Thomsen, and Jim Sidanius found that Black faculty were generally rated more highly than white faculty, in a study of 5,655 students at the University of Texas, and students placed more emphasis on the academic competence of Black than of White professors. ("Perceived Academic Competence and Overall Job Evaluations: Students' Evaluations of African American and European American Professors." *Journal of Applied Social Psychology* 39.2 [2009] 389–406). Other studies have focused on more specific characteristics, for instance Anderson and Smith found an interaction between women's ethnicity and teaching style in students' evaluations of their warmth, and on differential criteria used by students to rate professors with different ethnic identities (Anderson and Smith, "Student' Preconceptions," 184-201).
54. Kristin Anderson and Gabriel Smith. "Students' Preconceptions of Professors: Benefits and Barriers According to Ethnicity." *Hispanic Journal of Behavioral Science* 27 (2005): 184-201.
55. Heather Laube, Kelley Massoni, Joey Sprague, and Abby Ferber. "The Impact of Gender on the Evaluation of Teaching: What We Know and What We Can Do." *NWSA Journal* 19.3 (2007): 87-104.
56. In the last decade, several student ratings instruments have been designed to measure learning, such as the IDEA Student Ratings of Instruction system **www.theideacenter.org/node/5** retrieved July 20, 2010**).**