What does research say about standard student evaluations of teaching (SET)?

A summary prepared by Cynthia J. Jameson, University of Illinois at Chicago

1. ***The gender of the professor influences student evaluations.*** Studies which claim otherwise are flawed in various ways, for example (Feldman's 1992, 1993) [Feldman, K. A. "College students' views of male and female college teachers: Part I – Evidence from the social laboratory and experiments", *Research in Higher Education 33,* 3, 317-75 (1992); Feldman, K. A., "College Students' Views of Male and Female College Teachers. Part II Evidence from Students' Evaluations of Their Classroom Teachers", *Research in Higher Education*, 34, 151-211(1993).] by averaging correlations between gender and student rating of the professor that are quite disparate, and aggregating ratings from studies that vary by discipline, type of institution, and unit of analysis. Sociologists who specialize in gender are puzzled by conclusions that gender has no impact on teaching evaluations. Three decades of scholarship has shown that gender is a significant factor in shaping interactions, practices, and outcomes in every major realm of human social life. The broad and consistent message across all of the studies in evaluation of work and workers is that people discount women's skills and effort, are not comfortable with women in positions of power, and respond poorly to women who overstep their culturally assigned bounds. Why would the classroom be any different? Research carried out using controls clearly indicate that gender of the professor influences student evaluations. For example, [Ellyn Kaschak, "Sex bias in student evaluations of college professors," *Psychology of Women Quarterly,* 2, 235-242 (1978).]: 50 male and 50 female students were given a set of descriptions of the teaching methods and practices of professors in various specialties. In the forms received by half of the students (25 males and 25 females) the professors were given names of the opposite gender from the professors in the forms received by the other half of the students. Kaschak found that the male students were biased against those professors assigned female names, while the female students were not. From Sprague and Massoni [Joey Sprague and Kelley Massoni, "Student evaluations and gendered expectations: What we can't count can hurt us", *Sex Roles,* 53, 779-793 (2005)]:

> *'Psychological research on social cognition has explored gender-specific evaluation processes under the rubric of "shifting standards." Whenever people are called on to make a judgment, they do so in relation to some point of reference. When an evaluation concerns a behavior or attribute that resonates with race or gender stereotypes, these stereotypes influence the standard or context used to judge a particular member of the group (Biernat, 1995; Biernat & Kobrynowicz, 1997; Biernat & Manis, 1994; Biernat, Manis, & Nelson, 1991; Kobrynowicz & Biernat, 1997, 1998). Bennett (1982) pointed to a gendered shift in the standard that is applied to evaluate college instructors. She surveyed undergraduate students on how much personal attention they both expected and received from their instructors, as well as how they would rate their instructors on availability outside of class. Students expected and reported getting more personal time from women than from men, and yet were more likely to rate women instructors as not available enough. These students' reference point for "enough" availability clearly shifted to a higher order for women teachers (see also Burns- Glover & Veith, 1995).'*

[Biernat, M. (1995). The shifting standards model: Implications of stereotype accuracy for social judgment. In Y. T. Lee, L. Jussim, & C. McCauley (Eds.), *Stereotypes: Perspectives on accuracy and inaccuracy* (pp. 87–114). Washington, DC: American Psychological Association. Biernat, M., & Kobrynowicz, D. (1997). Gender- and race-based standards of competency: Lower minimum standards but higher ability standards for devalued groups. *Journal of Personality and Social Psychology*, *72,* 544–557. Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology*, *66,* 5–20. Biernat, M., Manis, M., & Nelson, T. E. (1991). Stereotypes and standards of judgment. *Journal of Personality and Social Psychology*, *60,* 485–499. Kobrynowicz, D., & Biernat, M. (1997). Decoding subjective evaluations: How stereotypes provide shifting standards. *Journal of*

Experimental Social Psychology, *33,* 579– 601. Bennett, S. K. (1982). Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology*, *74,* 170–179. Burns-Glover, A., & Veith, D. (1995). Revisiting gender and teaching evaluations: Sex *still* makes a difference. *Journal of Social Behavior and Personality*, *10*(6), 69–80.]

From Sprague and Massoni, 2005:

> *'S(s)tereotypes that people hold for a particular group can influence their understanding of the meaning of a trait in members of that group. In the case of teacher evaluations, students' gender stereotypes are apt to "shift" not only their baseline expectations for their teachers' traits, but also their perceptions of what those traits entail. Thus, students may expect a woman professor to engage in a different set of behaviors to satisfy a particular standard than they would expect of a man professor. For example, students may expect a woman professor to spend office hours walking them through a task when they might expect a man professor to only give brief directions in class. ... many shifts in the standards people apply take place without conscious awareness, and some cannot be stopped, even with effort (Biernat et al.,1991).'*

[Biernat, M., Manis, M., & Nelson, T. E. (1991). Stereotypes and standards of judgment. *Journal of Personality and Social Psychology*, *60,* 485–499.]

2. In typical student evaluations of teaching, the students are given a list of traits and behaviors and asked to rate their teacher on each using a numerical scale. The mean scores on each item are then taken as an indication of teaching performance. It has been pointed out [Heather Laube, Kelley Massoni, Joey Sprague, and Abby L. Ferber, "The Impact of Gender on the Evaluation of Teaching: What We Know and What We Can Do", *NWSA Journal* 19, 87-104(2007)] that ***evaluating teaching effectiveness using student evaluation questionnaires is based on two assumptions that the research literature suggests are untenable.*** First, it assumes a universal metric: that a "3" is a "3" and a "5" is a "5," no matter who the teacher is. Second, it assumes that a specific rating corresponds to equivalent behaviors or abilities across professors and instructors. But if, as the research suggests, students use different baselines for men and women, or, in some cases, they draw on totally different behaviors to evaluate a trait, then the scores on these student evaluations of teaching are not measures of the same things for female professors as for male professors.

Sprague and Massoni, 2008 conducted a study to discover whether undergraduate student evaluations of their teachers are influenced by gender by asking the students to think of the "best ever" teacher they have had and the "worst ever" teacher they have had and to describe these teachers using their own words. If there were no gender influences on evaluations, roughly the same incidences of roughly the same set of words would have been used by the students to describe their teachers. The results were as follows:

> *'...words that mean Arrogant, Uncaring, and Insensitive were used more often to describe men, whereas words that mean Unfair, Rigid, and Ignorant were used more often to describe women. Some synonym clusters were completely gender-specific, that is, some kinds of meaning were only used to describe either men or women worst teachers. Words that mean Out-of-touch and Pretentious were used to describe only men teachers; words that mean Bitch, Psychotic, and Unhappy were used to describe only women teachers.'*

> *'The cluster of words that mean RUDE is an example of a gender-loaded dimension in which the number, type, and range of words vary markedly with the gender of the teacher. For example, although students used the same number of words to describe RUDE male and female worst teachers (54 each), they utilized a greater diversity of*

words to describe RUDE women than to describe RUDE men (24 vs. 15). A comparison of the connotations of the clusters within the factor RUDE also helps to illuminate the gender differences. Although the cluster Hypocritical is somewhat gender-equivalent, the remaining three clusters of Angry, Loud, and Rude reveal gender disparities. Women teachers were described as Angry more often and with a greater range of words than were men teachers. Whereas men teachers were described as angry and grouchy, women teachers were described as angry, irritable, cranky, ill-tempered, snappy, temperamental, and as having a temper or an ill-temper. Conversely, men teachers were more likely to be described as Loud and related synonyms. Still, although one Loud man teacher was called a "smart alec," his Loud female counterpart was called a "smart ass," which perhaps again indicates the greater disrespect or even hostility leveled at bad women teachers. Finally, whereas men teachers were called Rude more often than women teachers, a greater diversity of words was used to describe Rude women, including "insolent," a word usually used to describe the behavior of a subordinate toward a superior.'

In other words, the student viewed the female teacher as a subordinate behaving insolently to her superior, a male student.

'... students described women teachers as MEAN twice as often as men teachers, and used three times as many words to do so. ...a much greater diversity of words are used to describe Mean women teachers. Although both worst men and women teachers were called mean, harsh, and nasty, only women teachers were called sarcastically mean, ill-willed, vindictive, threatening, inimical, and described as out to ridicule, fail, and punish students. Finally, perhaps the most obvious gender-loaded cluster in this factor is Bitch, which occurred for only women teachers and has no equivalent counterpart for men. The words used by students to describe their women teachers that are represented by this cluster—bitch, bitchy, bitch toward male students, witch, and feminazi—seem particularly angry and reveal very specific attacks on women teachers as women, rather than merely as bad teachers. No equivalently insulting and gender-specific slang terms were used to describe men teachers.'

3. Research documents that ***people who violate expectations generally are rated more negatively than people who behave as expected***. To receive good evaluations, male professors simply must demonstrate their competence and knowledge; that is, they need to fulfill their stereotypical gender role expectations. But female professors bear a double burden: they must fulfill both their gender role by being nurturant and warm, as well as their professional role by being competent and knowledgeable. Basow [Susan A. Basow, "Student Ratings of Professors are not Gender Blind" *AWM Newsletter*, Vol. 24, No. 5, Sept.-Oct. 1994] notes that in college/university teaching, males are the norm. Men are professors, women are women professors. Thus women are marked for gender in ways men are not. Researchers who consider the gender of the student rater find that the ratings of male professors are unaffected by student gender, that students appear to respond to male professors in a uniform manner, regardless of their own gender. Students respond differently to female professors, however, perhaps because women faculty are still a minority (this is most marked in STEM departments particularly). In one study by Basow and Silberg [Susan A. Basow, and Nancy T. Silberg "Student evaluation of college professors: Are female and male professors rated differently?" *Journal of Educational Psychology*. 79(3), 308-314 (1987).], 16 female professors were matched with a male professor in the same division, at the same rank, and with the same rank, and with the same number of years at the college. More than 1,000 students in classes taught by these 32 professors filled out two questionnaires. One was a standard student rating form consisting of 26 questions, summarized into five factor scores (scholarship, organization/clarity, instructor-group interaction, instructor-student interaction, and dynamism/enthusiasm) and an overall rating. The second (the Bern Sex Role Inventory) asked students to rate their professor on two sets of personality traits: instrumental (such as assertive or dominant), often viewed as

"masculine", and expressive (such as warm or nurturant), often considered "feminine". The results revealed a consistent pattern. On all five factor scores and the overall rating, male students rated female professors more negatively than they rated male professors - and generally more negatively than did female students in the same class. This type of interaction between the gender of the student and the gender of the professor has been found in laboratory research, but less frequently in field studies, which typically neglect to ask the gender of the student rater or fail to match professors on important variables like rank and discipline.

Basow and Silberg, 1987 ask: 'Why do male students tend to rate certain female faculty more poorly than male faculty? Male students may be more influenced by gender stereotypes than are female students.' Research has documented that men, compared to women, hold more traditional attitudes toward gender roles and demonstrate more bias against gender-role violators. In the Basow and Silberg study, males majoring in business and economics or in engineering rated female faculty most negatively. They found that those students have the most traditional attitudes toward women and gender roles. The effects of gender on student ratings of professors are complex but real, and should not be dismissed. These effects may be quite marked for specific teachers. For instance, a female teacher whose direct teaching style lacks marked warmth or friendliness may find the cards stacked against her when teaching male students in a field where women are a rarity.

For example, separate studies led by Sheila Bennett [Bennett, Sheila Kishler. "Student perceptions of and expectations for male and female instructors." *Journal of Educational Psychology,* 74, 170-179 (1982).] and Anne Statham [Statham, Anne, Richardson, Laurel, and Cook, Judith (1991). *Gender and University Teaching: A Negotiated Difference.* SUNY Press] found that women professors are judged more negatively than males if they are not more interested in and available to students than male professors. But even when women professors are more available and more helpful, their overall ratings are no higher. In order to receive comparable ratings, female professors need to do more than their male counterparts. Thus, findings of no difference between male and female professors in overall ratings may mask the fact that different standards are being used to judge male and female faculty.

4. *Nonverbal behaviors and personality characteristics significantly predict global end-of-semester student evaluations of teachers.* Research by Clayson and Sheffet [Dennis E. Clayson and Mary Jane Sheffet, "Personality and the student evaluation of teaching," *Journal of Marketing Education* 28, 149-160 (2006)] sought to find out (1) Does a relationship exist between personality characteristics and student evaluation of teaching in marketing and business core classes? The study found a consistent and positive relationship between course and instructor evaluations and personality measures. (2) If a relationship does exist between personality and evaluation, how early in the term does it develop? Their findings provide the answer: within the first five minutes of initial contact, during which, the students had not yet seen the course syllabus nor yet been exposed to any pedagogical interchange. Within fewer than 5 minutes of initial contact, students' perception of the personality of the instructor is associated with the final evaluation given 16 weeks later. Indeed, it has been shown in a separate study by Ambady and Rosenthal [Nalini Ambady and Robert Rosenthal, "Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness", *Journal of Personality and Social Psychology*. 64 (3), 431-441 (1993)] that judgments of college teachers' nonverbal behavior based on very brief (under 30 seconds) silent video clips significantly predicted global end-of-semester student evaluations of teachers. This and many examples of the influence of non-verbal behavior on student evaluations has been discussed in detail in a comprehensive review by Deborah Merritt [Deborah Meritt, "Bias, the brain, and student evaluations of teaching". *St. John's Law Review*, 82, 235-287 (2008.]
From Merritt:

> *'What role do nonverbal behaviors play in more routine student evaluations? Several researchers followed up on this question by using the Dr. Fox paradigm to conduct controlled classroom experiments. [See, e.g., Herbert W. Marsh & John E. Ware, Jr.,*

"Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: New Interpretations of the Dr. Fox effect", J. Educ. Psychol. 74, 126, 126–27 (1982) (reviewing earlier studies finding that students could be fooled into giving favorable evaluations of teachers when lectures are delivered in an enthusiastic and expressive manner);; Reed G. Williams & John E. Ware, Jr., "An extended visit with Dr. Fox: validity of student satisfaction with instruction ratings after repeated exposures to a lecturer", Am. Educ. Res. J. 14, 449, 449–50 (1977).] These studies used videos that systematically varied a lecturer's content and nonverbal behaviors to examine their relative effect on student teaching evaluations. A meta-analysis of this cluster of investigations concluded that nonverbal behaviors dramatically affected evaluations. For example, an entertaining style increased an instructor's ratings by about 1.2 points on a five point scale.[See Philip C. Abrami et al., "Educational seduction", Rev. Educ. Res. 52, 446, 455 (1982).] Lecturers who provided more content on the other hand received "inconsistent and generally much smaller" boosts in their evaluations. Other studies have isolated some of the specific nonverbal behaviors that generate positive student ratings. Based on a detailed analysis of university classes and student evaluations, Harry Murray determined that a professor's speech patterns, facial expressions, and humor had the greatest impact on student evaluations.[See Harry G. Murray, "Classroom teaching behaviors related to college teaching effectiveness", in Using Research To Improve Teaching (Janet G. Donald & Arthur M. Sullivan eds., 1985)] More learning-focused behaviors, such as giving "concrete examples of concepts," "point[ing] out practical applications," "repeat[ing] difficult ideas," or "providing sample exam questions" correlated less with student ratings. While the Fox studies suggested that faculty could reap greater evaluation rewards by focusing on style rather than substance, Murray's investigation sounded a further disturbing note: Even when concentrating on the stylistic elements of their teaching, faculty can more effectively raise student evaluations by using certain facial expressions than by offering concrete examples or repeating difficult concepts.'

5. The cumulative research suggests that there is ***little, if any, positive association between the ratings students give faculty and the amount they learn***. The most recent study suggests a negative correlation between evaluations and learning. In a particularly well-designed investigation, two business professors gathered 8 full years of data on students who completed two sequential accounting courses at a mid-western university. After controlling for factors (ACT scores, overall GPA, and grades in the first course) the researchers discovered that students who completed the first course with highly rated professors achieved significantly *lower* grades in the second course. [Penelope J. Yunker & James A. Yunker, Are Student Evaluations of Teaching Valid? Evidence from an Analytical Business Core Course, *J. Educ Bus*. **78**, 313–17 (2003).]

6. *What criteria are students using in filling out student evaluations of a class?*
A research study of students completing student questionnaires on teaching reveal what students say about these instruments.[Jordan J. Titus, Student Ratings in a Consumerist Academy: Leveraging Pedagogical Control and Authority *Sociological Perspectives*, Vol. **51**, Issue 2, pp. 397–422 (2008)] In this study, the categories were the usual ones found in most SET questionnaires. Rating sheets used had the following general categories (1) "the course as a whole," (2) "the course content," (3) "the instructor's contribution to the course," and (4) "the instructor's effectiveness in teaching the subject matter." Many students report that rather than reading the actual rating items, they locate a column on the form to reflect their general level of enjoyment in the course and then mark all of the rating items in that same column at that same value: "I find that it wastes my time, and it's boring, and anyways, the whole time I just fill in the fair or the good circle if I like the class, and I don't pay attention to the questions." In this way, students' enjoyment gains a distorted level of importance in SETs. Because their sense of

enjoyment is so widely used by students as the sole criterion by which they rate every item on the form, their level of pleasure becomes conflated with teaching quality. The ratings these students give are not considerations of specific teaching behaviors; instead, their ratings represent their general opinion of the instructor's acceptability. Students interviewed in the study do not differentiate between an "instructor's contribution to the course" and the "instructor's effectiveness in teaching the subject matter"; both are viewed by students as products of instructor likeability: "I would say personality would be the biggest contribution to a course an instructor could give, to make the course interesting by trying to entertain us while we're learning."

7. *In the face of the research partly described here, how can we continue to rely entirely upon conventional teaching evaluations to tell us what we want to know about a professor's teaching effectiveness?* Deborah Merritt notes in her in a comprehensive review (Deborah Meritt, 2008):

> *'Throughout the academy, faculty question whether student evaluations of teaching accurately reflect a professor's success in helping students learn. Many charge that evaluations actually undermine learning by encouraging lenient grading and superficial classroom presentations. [See, e.g., Valen E. Johnson, Grade inflation: a crisis in college education, pp. 235–37 (2003); Dennis E. Clayson & Mary Jane Sheffet, "Personality and the student evaluation of teaching", J. Marketing Educ. 28, 157–58 (2006); Charles R. Emery et al., "Return to academic standards: a critique of student evaluations of teaching effectiveness", Quality Assurance Educ. 11, 37–45 (2003); Wendy M. Williams & Stephen J. Ceci, "How'm I doing?," Change 29, 13–14 (1997)] In an increasingly diverse and competitive workplace, can we rely upon conventional teaching evaluations to tell us what we want to know about a professor's classroom success? Or do these evaluations reflect—and perhaps reinforce—biases based on race, sex, and other unwelcome characteristics?'*

Departments and administrators have an ethical and legal obligation not to base promotion and salary decisions on data which are biased.