The Drunkard's Search: Student Evaluation in Assessing Teaching Effectiveness

Christi Siver, Associate Professor, College of St Benedict | St John's University
Claire Haeg, Professor, College of St Benedict | St John's University

Abstract

As social scientists, we understand the problem of the Drunkard's Search — the lure and perils of using easy-to-obtain but irrelevant data — yet we are employed by institutions that are clearly searching under the lamppost for data to use in employment decisions. Researchers from various disciplines have studied and lamented the biases inherent in student course evaluations. Studies have found that these evaluations show systematic bias against women and people of color. They also may mask poor teaching practices as they are better measures of popularity than teaching effectiveness. Each year another study is released, leading to momentary hand wringing about the weakness of course evaluations as a means of assessing faculty and warning against their use in tenure and promotion decisions. Rather than adopting alternative means of evaluating faculty teaching and thinking creatively about student feedback, administrators and faculty leaders default to these surveys, claiming there is no other way to collect data on teaching. These course surveys continue to be used despite the mounting evidence that they provide not only no evidence of teaching effectiveness, but bad evidence that does damage to faculty both directly and indirectly. In an effort raise the profile of these discussions and push for tangible change, we offer a comprehensive literature review of the existing research on student course evaluations and their biases. While individual faculty may find some useful information in seeking feedback from students for their own teaching and course development, the review of survey findings outside of the context of the specific course and students offers little of value for department chairs or rank and tenure committees.

Junior faculty approaching the tenure application process and department chairs guiding candidates can use this study and its findings to properly contextualize course evaluations for their colleagues outside of the field. This study should also lead to conversations in departments, in faculty governance, and with university administrators about moving on from this easy but misleading system of evaluation and providing resources to provide greater faculty observation and mentoring as an indicator of teaching success.

Following is an annotated bibliography of the research on the various cognitive biases in student course evaluations and student course evaluations validity in measuring teaching effectiveness. Our goal is to create a comprehensive catalog of existing research on student course surveys to better track the development and sophistication of findings related to student course surveys and to guide discussions about potential alternatives or complementary approaches to assessing/evaluating teaching skill and student learning.

Anderson, Kathryn H., and John J. Siegfried. 1997. "Gender Differences in Rating the Teaching of Economics." *Eastern Economic Journal* 23 (3):347-357.

> Using data collected when revising the national test of economics students – the Test of Understanding College Economics, or TUCE – the authors studied 2408 2408 introduction to micro-economics students and 2185 introduction to macroeconomics students, and examined the data for evidence of teaching quality and gender bias in instructor evaluations. The study examined a wide variety of variables, including student and instructor race, gender, ethnicity, instructor accent and experience and class size. The authors find no evidence that instructor gender has a negative impact upon teacher quality or evaluation. They find that women students have significantly less interest in economics than their male peers.

Arbuckle, Julianne, and Benne D. Williams. 2003. "Students' Perceptions of Expressiveness: Age and Gender Effects on Teacher Evaluations." *Sex Roles* 49 (9/10):507-516.

> In a large-N experiment, the authors use an experiment where they asked 352 undergraduate students to watch slides of a gender- and age-neutral stick figure while listening to a neutral voice presenting a lecture. Students were then asked to rate the lecture on forms that indicated the age and gender of the instructor. The authors find that students rated "young male" instructors higher than "young female" "old male" or "old female" professors on speaking enthusiastically and using a meaningful voice tone, even though the lecture was exactly the same in each case.

Babcock, Linda, Maria P. Recalde, Lise Vesterlund, and Laurie Weingart. 2017. "Gender Differences in Accepting and Receiving Requests for Tasks with Low Promotability." *American Economic Review* 107 (3):714-747.

> Using an experimental design, Babcock et al test whether women are altruistic and thus are more likely to volunteer for tasks with "low promotability." The authors distinguish between high-promotability tasks, like research, and low-promotability tasks, like university service. The authors use a variety of experiments to test whether gender impacts an individual's willingness to volunteer for tasks that everyone sees as beneficial, but that they wish someone else would do. They find, over several different tests, that women do not differ from men in terms of altruism or risk. Instead, gender perceptions lead women to take on these undesirable tasks. They note "Both men and women are less likely to volunteer when

there are more women in the group. This response to gender composition is consistent with the belief that women more than men will volunteer" (743). They note that accepting low-promotability tasks may slow women down in their careers and can snowball if they are repeatedly asked to take on such tasks. The authors suggest that the assignment of such tasks should not be discretionary; they should be more equally allocated among faculty.

Baldwin, Tamara, and Nancy Blattner. 2003. "Guarding against Potential Bias in Student Evaluations: What Every Faculty Member Needs to Know." *College Teaching* 51 (1):27-32.

An instructive article that discusses both the major issues of bias in SETs and potential alternatives to SETs in the evaluation of teaching. Alternatives include 1) using a variety of teaching evaluation methods, including adding instructor- and technique-specific questions to the general evaluation; 2) administer multiple evaluations throughout the course; 3) create a teaching portfolio, including syllabi, assignments, examples of student feedback, etc.; 4) inviting peer observation; 5) collecting formal and informal feedback from students; and 6) putting student comments in context.

Basow, S. A. 1995. "Student evaluations of college professors: When gender matters." *Journal of Educational Psychology* 87 (4):656-665. doi: 10.1037/0022-0663.87.4.656.

In a study of SETs gathered over four semesters at a small liberal arts college, the authors find significant teacher gender/student gender interaction. The effect sizes are small, however. Male professors' evaluations are unaffected by student gender, whereas women professors were more highly rated by female students.

Basow, S. A., J. E. Phelan, and L. Capotosto. 2006. "Gender patterns in college students' choices of their best and worst professors." *Psychology of Women Quarterly* 30 (1):25-35.

Using an open-ended questionnaire, 175 students were asked to describe their "best" and "worst" professors. The authors find that gender dynamics were most evident in the pairing of male students with female professors. Male students were most critical of their female professors and based their criticisms on poor classroom dynamics, particularly perceived close-mindedness of women professors.

Basow, Susan A., Stephanie Codos, and Julie L. Martin. 2013. "The effects of professors' race and gender on student evaluations and performance." *College Student Journal* 47:352+.

Basow et al focus on the potential impact of gender and race bias in SET. Noting the importance of SET in higher education hiring and promotion decisions, they find that while some studies have discovered systematic gender bias in evaluations, much less research has been done on the role of race. Using an experimental design during which students watched a virtual online lecture by a variety of different idealized professors (a white man, a white woman, an African American man, and an African American man). While the evaluations of the "professors" were not consistent with the expectations of gender and race bias, the subsequent tests of student learning suggested that students learned more from the white male instructor than from the white woman or either African American professor. They

argue that the evaluations may be tainted by student notions of "correct" social responses regarding gender, the subsequent learning quizzes may demonstrate that students pay more or less attention to information delivered based on their underlying biases.

Batten, J., P. D. J. Birch, J. Wright, A. J. Manley, and M. J. Smith. 2014. "An exploratory investigation examining male and female students' initial impressions and expectancies of lecturers." *Teaching in Higher Education* 19 (2):113-125. doi: 10.1080/13562517.2013.827645.

The authors surveyed 752 students who rated 30 informational cues in terms of the extent to which these cues influence the students' initial impressions of their lecturers. Students report that factors such as instructor voice clarity and control of class are more important to them than instructor gender, race, or nationality.

Beitzel, Brian D. 2013. "Student Response to Faculty Instruction (SRFI): An Empirically Derived Instrument to Measure Student Evaluations of Teaching." *Journal of Research in Education* 23 (2):97-115.

Describes the development of the SRFI instrument.

Birch, P. D. J., J. Batten, A. J. Manley, and M. J. Smith. 2012. "An exploratory investigation examining the cues that students use to form initial impressions and expectancies of lecturers." *Teaching in Higher Education* 17 (6):660-672. doi: 10.1080/13562517.2012.658561.

The authors fielded a questionnaire that asked respondents to report on their initial impressions of faculty instructors and determine what the most important factors were. The study examined 452 student responses and determined that dynamic (non-demographic) rather than static (race, gender, the wearing of eyeglasses) cues were more important. That authors observe that such self reporting might miss student gender expectations and bias.

Boswell Stefanie, S. 2016. "Ratemyprofessors is hogwash (but I care): Effects of Ratemyprofessors and university-administered teaching evaluations on professors." *Computers in Human Behavior* 56:155-162.

Boswell conducts a comparison of student evaluations of teaching on the website "RateMyProfessor.com" and traditional university-run SETs using evaluations of the same course. She concludes that RateMyProfessor contains "useful" feedback but that professors pay less attention to the site than they do their own SETs.

Braga M, Paccagnella M, and Pellizzari M. 2014. "Evaluating Students' Evaluations of Professors." *Economics of Education Review* 41: 71–88. doi:10.1016/j.econedurev.2014.04.002.

Braga and Pellizzari explore the assumed connection between student course evaluations and student learning. While institutions currently use student course evaluations to evaluate teachers based on the assumption that they serve as measure of student learning, the authors find "teachers who are associated with better subsequent performance receive worse evaluations from their students" (72). Using a natural experiment where students are

randomly assigned to instructors for their compulsory courses, the authors track the students through their academic careers. By comparing student performance and student evaluations, they find that student learning did not correlate with positive student evaluations. They find "teachers who are more effective in promoting future performance receive worse evaluations from their students. This relationship is statistically significant for all items … and is of sizable magnitude" (81). They also note that irrelevant external conditions, like the weather, can influence student course evaluations (84). They offer some possible suggestions to improve the validity of student course evaluations; better students' evaluations could be weighted more heavily, students could be surveyed at a later point in their academic career, or the institution could implement other evaluation mechanisms, such as faculty observations. They conclude "these measurement modes – as well as other potential alternative[s] are costly, but they should be compared with the costs of the current systems of collecting students' opinions about teachers which are often non-trivial" (86).

Braidwood, T., and J. Ausderan. 2017. "Professor Favorability and Student Perceptions of Professor Ideology." *PS-Political Science & Politics* 50 (2):565-570. doi: 10.1017/s1049096516003206.

Using a specially designed survey instrument fielded to 332 students (62 political science majors) the authors measure the effect of student ideology and professor favorability (whether the student liked the professor or not) on the student's assessment of the professor's ideology. The authors find that students project their own ideology on professors that they like, and an opposite ideology on professors that they dislike.

Burns-Glover, A. L., and D. J. Veith. 1995. "Revisiting gender and teaching evaluations: Sex still makes a difference." *Journal of Social Behavior and Personality* 10 (6):69-80.

The authors developed a list of sex-stereotyped traits as well as a student generated list of desirable traits in a professor, and then asked 75 respondents to rank "Sam," "Sarah," and "Dr." Lawson – a supposed applicant for a university teaching job. Students rated professors differently depending on the professor's supposed gender. Students presumed that the professor was male when the target was labelled as "Dr." Female professors were expected to be available outside of class time, to be flexible and open to change, and to care about students.

Butcher, Kristin F., Patrick J. McEwan, and Akila Weerapana. 2014. "The Effects of an Anti-Grade-Inflation Policy at Wellesley College." *Journal of Economic Perspectives* 28 (3):189-204. doi: 10.1257/jep.28.3.189.

A natural experiment at Wellesley College in which department average GPA in introductory courses was restrict. The resulting decreases in instructor evaluation scores indicates that SETs are impacted by student expected grade.

Carrell, Scott E., Marianne E. Page, and James E. West. 2010. "Sex and Science: How Professor Gender Perpetuates the Gender Gap." *The Quarterly Journal of Economics* 125 (3):1101-1144.

Through a study of the (random) assignment of students at the US Air Force Academy to introductory classes, and those students' subsequent choices of major, the authors find that while professor gender has little impact on male students, it has a significant impact of female students' performance in math and science courses, as well as the likelihood of their taking future math and science courses and becoming STEM majors.

Carrell, Scott E., and James West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *Journal of Political Economy* 118 (3):409-432. doi: 10.1086/653808.

The authors examine the grades and SETs of students in introductory calculus courses and compare them with the students' grades in subsequent mathematics courses. Students rewarded high grades in the introductory course with high ratings on SETs, but punished professors who provided challenging teaching environments (and thus "deep learning") yet the students who had these challenging (low-SET score) professors did better in subsequent courses.

Carter, Robert E. 2016. "Faculty Scholarship Has a Profound Positive Association with Student Evaluations of Teaching--Except When It Doesn't." *Journal of Marketing Education* 38 (1):18-36.

Using a statistical analysis of various datasets (including RateMyProfessor) the author examines 2720 evaluations for 219 faculty across 21 universities over 10 years. The author focuses on the impact of faculty scholarship on teaching evaluations, and finds that only male faculty at flagship universities with an elite publication obtain higher SETs than faculty without this publication record at non-flagship schools

Centra, John A., and Noreen B. Gaubatz. 2000. "Is There Gender Bias in Student Evaluations of Teaching?" *The Journal of Higher Education* 71 (1):17-33. doi: 10.2307/2649280.

Student course surveys from 741 classes at various institutions (including 2 year and 4 year) were examined for gender bias. Each class had at least 10 female and 10 male students. Female instructors received lower scores from male students and higher from female students. Women instructors were more likely to use discussion techniques. The differences, although statistically significant, were not large.

Costin, Frank, William T. Greenough, and Robert J. Menges. 1973. "Student Ratings of College Teaching: Reliability, Validity, and Usefulness." *The Journal of Economic Education* 5 (1):51-53. doi: 10.2307/1182836.

An early literature review on the scholarship of SETs. The authors find that SETs could provide valuable and valid evaluations of teaching. The authors note that, where other types of evaluation of teaching (such as peer evaluation) is available, SETs show low correlation, and thus might provide a valuable corollary.

dApollonia, S., and P. C. Abrami. 1997. "Navigating student ratings of instruction." *American Psychologist* 52 (11):1198-1208. doi: 10.1037//0003-066x.52.11.1198.

In an field review of multi-section empirical studies of student assessment of teachers, the authors find that SETs have validity, but are nevertheless affected by an instructor's rank, experience, and autonomy, course discipline and class size.

Das, M., and H. Das. 2001. "Business students' perceptions of best university professors: Does gender role matter?" *Sex Roles* 45 (9-10):665-676. doi: 10.1023/a:1014867809922.

The authors surveyed 292 business schools students at two Canadian univeristies about their "best professor." Student respondents were asked to fill out a questionnaire regarding their own demographics and then fill our a Sex Role Inventory about their "best" professor. The authors find that "a student's own gender and gender role are significantly related to those of his/her best professor"

Davidovitch, Nitza, and Dan Soen. 2009. "Myths and Facts about Student Surveys of Teaching the Links between Students' Evaluations of Faculty and Course Grades." *Journal of College Teaching & Learning* 6 (7):41-50.

The authors conduct an analysis of a large dataset of instructor SET scores (N=16,484) using five different regression models. This study found no correlation between student grades and instructor ratings; the authors did find correlations between student evaluations fo the course's structure and design and the instructor's age and tenure.

De Witte, K., and N. Rogge. 2011. "Accounting for exogenous influences in performance evaluations of teachers." *Economics of Education Review* 30 (4):641-653. doi: 10.1016/j.econedurev.2011.02.002.

Using a complex statistical analysis of SETs in 112 college courses taught by 69 different faculty. The authors find that teacher gender does not significantly influence evaluations. One of the models shows that instructor age might.

De Witte, K., N. Rogge, L. Cherchye, and T. Van Puyenbroeck. 2013. "Accounting for economies of scope in performance evaluations of university professors." *Journal of the Operational Research Society* 64 (11):1595-1606. doi: 10.1057/jors.2012.115.

Using complex statistical modelling of SETs of professors in a large business school, the authors find that scholarship decreases SET ratings of professors.

Dukes, Richard L., and Gay Victoria. 1989. "The Effects of Gender, Status, and Effective Teaching on the Evaluation of College Instruction." *Teaching Sociology* 17 (4):447-457. doi: 10.2307/1318422.

In a study of 144 undergraduate students, each student was given a packet describing four scenarios of a teacher in a college class. The scenarios varied in terms of four variables: knowledge, enthusiasm, rapport, and organization of instructor. The authors found no significant impact of instructor gender on evaluations.

Farreras, Ingrid G., and Robert W. Boyle. 2012. "The Effect of Faculty Self-Promotion on Student Evaluations of Teaching." *College Student Journal* 46 (2):314-322.

In a quasi-experimental study, 322 students enrolled in psych courses (of whom 80% were women) were provided with packets that outlined the biography of a lecturer along with a transcribed lecture. The lectures were varied to include four different types of self-promotion by the lecturer. The biographies differed in terms of the speaker's gender. Although some students did perceive some self-promotion and punished the speaker for boasting, they did not see self-promotion in all the ways the researchers did. There was no perceived gender difference, perhaps because it was a transcript of a lecture.

Feldman, Kenneth A. 1992. "College Students' Views of Male and Female College Teachers: Part I: Evidence from the Social Laboratory and Experiments." *Research in Higher Education* 33 (3):317-375.

A review of the scholarly literature of gender bias in SETs, focusing on experimental and laboratory studies.

Gentry, Jeffery. 2011. "Radical Change in Faculty and Student Evaluation: A Justifiable Heresy?" Administrative Issues Journal: Education, Practice, and Research 1 (1):57-64.

A normative discussion of the field. Gentry argues that SETs give students a means of punishing rigorous instructors.

Goos, Maarten, and Anna Salomons. 2017. "Measuring teaching quality in higher education: assessing selection bias in course evaluations." Research in Higher Education 58 (4):341-364. doi: 10.1007/s11162-016-9429-8.

A large-N study of online SETs from 28K students taught across 3000 courses, examining the impact of non-response-rates on SET validity. The authors find that low response rates actually increase instructor approval.

Hameed, F., A. Ali, A. Hameed, Z. Saleem, and Y. Javed. 2015. "Teacher evaluation: the role of gender." *Quality & Quantity* 49 (5):1779-1789. doi: 10.1007/s11135-014-0054-3.

The authors fielded a structured questionnaire to 250 students from four different universities in Pakistan and found that instructors were evaluated equally by the male and female students. In addition, the authors concluded that instructors showed some gender bias in involving their students in classroom activities and recognizing their efforts

Hamermesh, Daniel S. and Amy Parker. 2005. "Beauty in the classroom: instructors' pulchritude and putative pedgagogical productivity." *Economics of Education Review* 24 (2005): 369-376.

Hamermesh and Parker ask whether student perceptions of beauty have an influence on student course evaluations. They note that teaching quality can have important impacts on earning potential for instructors. Drawing on the evaluations of instructors at various levels in terms of academic hierarchy, the authors assembled data on 463 classes. Instructor pictures were rated by undergraduate students and then compared with the course

evaluations for that instructor. Beauty had a significant impact on the evaluations; they note "Moving from one standard deviation below the mean to one standard deviation above leads to an increase in the average class rating of 0.46, close to a one-standard deviation increase in the average class rating" (372). The impact of beauty doubles in lower-division courses. The authors warn "even if instructional ratings have little or nothing to do with actual teaching productivity, university administrators behave as if they believe that they do, and they link economic rewards to them" (375). Given the arbitrariness the authors demonstrate inherent in student course evaluations, chairs and evaluators should be extremely cautious in using course evaluations as evidence of teaching effectiveness.

Harris, Mary B. 1975. "Sex Role Sterotypes and Teacher Evaluations." *Journal of Educational Psychology* 67 (6): 751-756.

Harris notes that while some previous research has not found gender-based differences in teaching evaluations, that research may not have been comparing equivalent male and female teachers. Alternatively, she suggests asking students to rate general descriptions based only on differences in gender. Using an experimental design, Harris had students answer questions based on booklets of information that different only in terms of the sex of the teacher, name, and pronouns. She found that the primarily difference in terms of the ratings was in terms of masculinity and femininity. Some males might be considered more feminine if they taught in a subject that was perceive as feminine (nursing). However, she also notes that "a teacher who used a feminine teaching style was rated less positively on all variables except warmth than a teacher who used a more masculine mode (755). She concludes that there was no evidence of a perception of discrimination against female professors.

Laird Thomas, F. Nelson, Amy K. Garver, and Amanda Suniti Niskode-Dossett. 2011. "Gender Gaps in Collegiate Teaching Style: Variations by Course Characteristics." *Research in Higher Education* 52 (3):261-277.

Using a statistical analysis of a survey of over 9,000 faculty members, the authors found that women spend less time lecturing and more time in active classroom practices. The gender gap is dependent upon course level and the number of of times instructor taught course in a prior semester.

Linse, Angela R. 2017. "Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees." *Studies in Educational Evaluation* 54:94-106. doi: 10.1016/j.stueduc.2016.12.004.

Provides a relatively extensive literature review of the scholarship of SETs organized to assist administrators and committees who use them.

Lucal, Betsy, Cheryl Albers, Jeanne Ballantine, Jodi Burmeister-May, Jeffrey Chin, Sharon Dettmer, and Sharon Larson. 2003. "Faculty Assessment and the Scholarship of Teaching and Learning: Knowledge Available/Knowledge Needed." *Teaching Sociology* 31 (2):146-161. doi: 10.2307/3211305.

An extensive literature review essay on the scholarship of student evaluations of teaching in regards to faculty assessment.

MacNell, Lillian, Adam Driscoll, and Andrea N. Hunt. 2015. "What's in a Name: Exposing Gender Bias in Student Ratings of Teaching." *Innovative Higher Education* 40 (4):291-303.

Using an experimental study within an online course, teaching instructors (one man and one woman) were each given two sections which each taught under two different gender identities. Students rated the male identity significantly higher than the female identity regardless of the actual gender of the instructor.

Marsh, Herbert W. 1991a. "Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures." *Journal of Educational Psychology* 83 (2):285-296. doi: 10.1037/0022-0663.83.2.285.

Marsh examines the existing literature on student course evaluations, noting the promising uses of such evaluations and the underdeveloped nature of existing research. He notes the importance of dimensionality in terms of providing greater reliability of measurement: multiple factors of evaluation provide more valid and reliable results than ad hoc or aggregate assessments. He notes that evaluations have to be carefully designed and weighted; most current surveys lack developed factor analysis. He also addresses concerns about student course evaluations. Drawing in his own and others' research, he notes that surveys are quite often well correlated to common measures of student learning and consistent with intructors' own evaluations of their courses. He raises doubts about the validity of peer observation, but is optimistic about the possibility of bringing in trained observers. Seeking out areas of systemic bias, he considers factors identified by instructors, including course difficulty, grading leniency, and instructor popularity (notably not gender, race, ethnicity, religion). He found that these factors could not explain residual variance in course evaluations.

Marsh, Herbert W. 1991b. "A multidimensional perspective on students' evaluations of teaching effectiveness: Reply to Abrami and D'Apollonia (1991)." *Journal of Educational Psychology* 83 (3):416-421. doi: 10.1037/0022-0663.83.3.416.

Martin, Lisa L. 2016. "Gender, Teaching Evaluations, and Professional Success in Political Science." *PS-Political Science & Politics* 49 (2):313-319. doi: 10.1017/s1049096516000275.

Using a statistical analysis of a huge publicly available SET datasets from two political science departments in large U.S. universities. The author finds that female professors get substantively and significantly lower evaluations in large courses because of role incongruity.

Mengel, Friederike, Jan Sauerman, and Ulf Zolitz. 2017. *Gender Bias in Teaching Evaluations*. Bonn, Germany: Institute on Behavior and Inequality.

Using a large quasi-experimental dataset that included SETs from over 19,000 evaluations from the Maastrict University business school. The authors found that women instructors receive systematically lower teaching evaluations, and that this finding was particularly pronounced for junior women.

Miller, J., and M. Chamberlin. 2000. "Women are teachers, men are professors: A study of student perceptions." *Teaching Sociology* 28 (4):283-298.

Researched conducted a survey of sociology students at a research university, asking them to identify the highest educational level attained by instructors within the department. Students tended to think that male instructors have a higher level of educational attainment than they really do, while female instructors have a lower level of education than they do in reality. Women are given the label of "teacher" while men are thought of as "professor" regardless of position or credential.

Morgan, Helen K., Joel A. Purkiss, Annie C. Porter, Monica L. Lypson, Sally A. Santen, Jennifer G. Christner, Cyril M. Grum, and Maya M. Hammoud. 2016. "Student Evaluation of Faculty Physicians: Gender Differences in Teaching Evaluations." *Journal of Womens Health* 25 (5):453-456. doi: 10.1089/jwh.2015.5475.

The authors examined over 14,000 teaching evaluations of 916 professors of medicine engaged in teaching clinical rotations. They found that female professors received lower evaluations in all four types of clinical rotations examined.

Nadler, Joel T., Seth A. Berry, and Margaret S. Stockdale. 2013. "Familiarity and sex based stereotypes on instant impressions of male and female faculty." *Social Psychology of Education* 16 (3):517-539. doi: 10.1007/s11218-013-9217-7.

In two different quasi-experimental studies, 105 psychology undergraduate students and law students were provided with photographs of familiar and unfamiliar professors. Among psychology students, familiarity increase the sex-bias against female faculty, whereas among law students familiarity decreased sex-bias against female faculty. The authors conclude that some sex-bias against instructors may be discipline dependent.

Nasser, F., and K. A. Hagtvet. 2006. "Multilevel analysis of the effects of student and instructor/course characteristics on student ratings." *Research in Higher Education* 47 (5):559-590.

Using a statistical examination of SETs from 1867 students in 117 courses, the authors examine endogenous variable impacts on SET scores for instructors. The authors find that students' expected grade, their interest in the course, and age had significant positive impact of instructor ratings. Instructor workload (full-time versus part-time) was correlated with lower ratings, and having more female students also leads to higher ratings.

Parks-Stamm, Elizabeth J., and Chanda Grey. 2016. "Evaluating Engagement Online Penalties for Low-Participating Female Instructors in Gender-Balanced Academic Domains." *Social Psychology* 47 (5):281-287. doi: 10.1027/1864-9335/a000277.

In a large study of 5375 ratings of instructors in 360 online courses, the authors find that women instructors are penalized for not participating and that male instructors are not.

Pittman, Chavella T. 2010. "Race and Gender Oppression in the Classroom: The Experiences of Women Faculty of Color with White Male Students." *Teaching Sociology* 38 (3):183-196.

In 17 in-depth interviews of women faculty of color at a predominantly white research university, the authors find that the faculty members in their student encountered a significant level of gendered racism in the classroom.

Potvin, G., and Z. Hazari. 2016. "Student evaluations of physics teachers: On the stability and persistence of gender bias." *Physical Review Physics Education Research* 12 (2). doi: 10.1103/PhysRevPhysEducRes.12.020107.

In a large survey (N=6772) of college students, the authors find that both male and female students under-rated female high school teachers of physics. Students with strong physics identity showed a strong bias in favor of male teachers. [Note: an earlier study of students found that male students underrated their female high school teachers of biology and chemistry.] The authors conclude that biased evaluations of feedback may be a problematic for women in science.

Price, Linda, Ingrid Svensson, Jonas Borell, and John T. E. Richardson. 2017. "The Role of Gender in Students' Ratings of Teaching Quality in Computer Science and Environmental Engineering." *IEE Transactions on Education* 60 (4):281-287.

In a large-N study of SETs at an engineering oriented school, the authors find that instructors tended to get higher ratings when they were teaching subjects that were not typed for their gender. Male and female instructors received different ratings, and there were differences in ratings depending on the gender of the student, but the impact was small.

Pritchard, Robert E., and Gregory C. Potter. 2011. "Adverse Changes in Faculty Behavior Resulting from Use of Student Evaluations of Teaching: A Case Study." Journal of College Teaching & Learning 8 (1):1-7.

In an examination of graduating student data gathered in SETs from a course every two years from 1998 to 2006, the authors conclude that professors relying on good SET results reduced the number of hours of work required in their courses in order to improve their SETs.

Risquez, A., E. Vaughan, and M. Murphy. 2015. "Online student evaluations of teaching: what are we sacrificing for the affordances of technology?" Assessment & Evaluation in Higher Education 40 (1):120-134. doi: 10.1080/02602938.2014.890695.

In a large scale and longitudinal study, the authors compare in-class versus online SETs and conclude that the administration method does not meaningfully impact the results.

Rosen, Andrew S. 2018. "Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of RateMyProfessors.com data." Assessment & Evaluation in Higher Education 43 (1):31-44. doi: 10.1080/02602938.2016.1276155.

In a large statistical analysis of the data in RateMyProfessor.com, the study examines almost 8 million student reviews of professors and finds correlations between high ratings and instructor attractiveness and high ratings and easiness. Science instructor ratings are lower than humanities ratings. The author also finds that women faculty have lower scores in some fields (history and political science, for example) but not in others (chemistry.) There are no disciplines where women have overall higher ratings.

Sinclair, L., and Z. Kunda. 2000. "Motivated stereotyping of women: She's fine if she praised me but incompetent if she criticized me." *Personality and Social Psychology Bulletin* 26 (11):1329-1342. doi: 10.1177/0146167200263002.

In relative small study (N=180) of students who were given an in-depth survey about the course soon after they had received a grade for the course. Participants viewed women instructors as less competent than men instructors after receiving poor grades from them. The authors link this behavior to gender-stereotyping behavior and point out that disparaging a harsh female evaluator permitted participants to challenge the self-evaluative implications of negative feedback

Spooren, Pieter, Bert Brockx, and Dimitri Mortelmans. 2013. "On the Validity of Student Evaluation of Teaching: The State of the Art." *Review of Educational Research* 83 (4):598-642.

In an extensive review of the literature on the scholarship of SET, the authors find that SETs alone do not cause better teaching. They also conclude that most studies are not generalizable because they are limited to a particular instrument used in a particular setting.

Spooren, P., and W. Christiaens. 2017. "I liked your course because I believe in (the power of) student evaluations of teaching (SET). Students' perceptions of a teaching evaluation process and their relationships with SET scores." *Studies in Educational Evaluation* 54:43-49. doi: 10.1016/j.stueduc.2016.12.003.

Spooren, Pieter, Frederic Vandermoere, Raf Vanderstraeten, and Koen Pepermans. 2017. "Exploring high impact scholarship in research on student's evaluation of teaching (SET)." *Educational Research Review* 22:129-141. doi: 10.1016/j.edurev.2017.09.001.

An analysis of the most cited and highest impact articles in the literature about SET. The authors identify Herbert Marsh as the most-cited author, but agree that much of the literature dates back to the 1960s and 1970s.

Sprague, J., and K. Massoni. 2005. "Student evaluations and gendered expectations: What we can't count can hurt us." *Sex Roles* 53 (11-12):779-793. doi: 10.1007/s11199-005-8292-4.

While most quantitative studies find no evidence of gender bias in teaching evaluations, sociologists who study gender question these finds. The authors undertake a qualitative (content analysis) study of 200 students' descriptions of their "best" and "worst" teachers. The authors find that students to some extent, have gender-specific expectations of teachers.

Stroebe Wolfgang, W. "Why Good Teaching Evaluations May Reward Bad Teaching: On Grade Inflation and Other Unintended Consequences of Student Evaluations." *Perspectives on Psychological Science* 11 (6):800-816.

A field review essay of SET scholarship focused on the impact of student grade expectations on student course evaluations. Most studies find this effect.

Thawabieh, Ahmad M. 2017. "Students Evaluation of Faculty." *International Education Studies* 10 (2):35-43.

A study of 5291 student evaluations of courses across disciplines. The study finds that students' gender, expected grades, and the discipline of the course have an impact on student course evaluations.

Tobin, R. G. 2017. "Too Early for Physics? Effect of Class Meeting Time on Student Evaluations of Teaching in Introductory Physics." Physics Teacher 55 (5):276-279. doi: 10.1119/1.4981033.

Wilson, Deborah, and Kenneth O. Doyle. 1976. "Student Ratings of Instruction: Student and Instructor Sex Interactions." The Journal of Higher Education 47 (4):465-470. doi: 10.2307/1978730.

Using a student course questionnaire, the authors completed a statistical analysis of 312 students from 12 fine arts, humanities and social science courses (six taught by men and six by women,) They found no evidence that instructor gender impacted course evaluations, nor did they find interaction effects of instructor gender and student gender.

Wright, S. K. 2013. "Instructors' Address Forms Influence Course Ratings." Names-a Journal of Onomastics 61 (2):92-100.

Using a quasi-experimental method in which 70 particpants were asked to rate a syllabus description that including information about the professor, the authors found that courses with the instructor labeled with an academictitle (i.e., Professor, Dr) received higher ratings than those with a generic title (i.e., Mr, Ms, Mrs, Miss), and those with a male address form received higher ratings than those with a female address form.

Wright, Stephen L., and Michael A. Jenkins-Guarnieri. 2012. "Student evaluations of teaching: combining the meta-analyses and demonstrating further evidence for effective use." Assessment & Evaluation in Higher Education 37 (6):683-699. doi: 10.1080/02602938.2011.563279.

An overview of prior research that examines 11 identified meta-analyses of research into the effectiveness of SETs. The authors find that SETs are valid and are mostly free from gender bias, although they do admit that only one of the meta-analyses was focused on gender and that later studies indicate that further research is required.

Zabaleta, F. 2007. "The use and misuse of student evaluations of teaching." Teaching in Higher Education 12 (1):55-76. doi: 10.1080/13562510601102131.

In a large-N study of SETs in a language program, the authors find that women instructors assigned higher grades and received slightly better evaluations, but the effect was small. There is a correlation between low grades and low evaluations, but no correlation between high grades and high evaluations. Gender was not a major focus of the study, and the was focused on a gender-normed field.

Zhu, Luke, Karl Aquino, and Abhijeet K. Vadera. 2016. "What Makes Professors Appear Credible: The Effect of Demographic Characteristics and Ideological Beliefs." Journal of Applied Psychology 101 (6):862-880. doi: 10.1037/apl0000095.

Using a quasi-experimental method and five separate studies, the authors find that a researcher's perceived status (gender, education, race, caste) impacts the researcher's credibility, but does so differently depending on the reader's egalitarian or elitist ideology.