Journal of Academic Ethics (2022) 20:73–84
https://doi.org/10.1007/s10805-021-09400-w

Evaluating Student Evaluations of Teaching: a Review of Measurement and Equity Bias in SETs and Recommendations for Ethical Reform

Rebecca J. Kreitzer[1] · **Jennie Sweet-Cushman**[2]

Accepted: 27 January 2021 / Published online: 9 February 2021

Abstract
Student evaluations of teaching are ubiquitous in the academe as a metric for assessing teaching and frequently used in critical personnel decisions. Yet, there is ample evidence documenting both measurement and equity bias in these assessments. Student Evaluations of Teaching (SETs) have low or no correlation with learning. Furthermore, scholars using different data and different methodologies routinely find that women faculty, faculty of color, and other marginalized groups are subject to a disadvantage in SETs. Extant research on bias on teaching evaluations tend to review only the aspect of the literature most pertinent to that study. In this paper, we review a novel dataset of over 100 articles on bias in student evaluations of teaching and provide a nuanced review of this broad but established literature. We find that women and other marginalized groups do face significant biases in standard evaluations of teaching – however, the effect of gender is conditional upon other factors. We conclude with recommendations for the judicious use of SETs and avenues for future research.

Keywords Teaching evaluations · Gender stereotypes · Gender bias · Gender

A contributing factor to disparities in tenure, promotion, and pay decisions is the near-ubiquitous reliance on student evaluations of teaching (SETs) (Murray, 1984; Wachtel, 1998). This dependency exists across institutions despite the abundance of research demonstrating ways in which student-driven evaluation can be problematic (Spooren et al., 2013; Stark & Freishtat, 2014; Wachtel, 1998), particularly for women and other marginalized groups who are fewer in number as rank increases[1]. In some disciplines, these

[1] For now, the entirety of this discussion and related research is binary in its orientation. We recognize that gender is more complex than women and men and acknowledge that gender identity that does not overtly conform to the binary likely complicates evaluations of teaching further than the existing body of knowledge has even identified

✉ Jennie Sweet-Cushman
jsweetcushman@chatham.edu

[1] University of North Carolina At Chapel Hill, Chapel Hill, North Carolina, USA

[2] Chatham University, Pittsburgh, PA, USA

disparities have been documented at each rank, from graduate student (APSA, 2011)—to full professor. Those that remain in the academy frequently face pay inequities.

While there are numerous considerations that contribute to these disparities (Aguirre Jr, 2000; Barbezat & Hughes, 2005; Perna, 2005; Bos et al., 2019), potential bias in SETs is particularly worrisome (Wagner et al., 2016; Murray, 1984). Criticism of SETs is a common topic of discussion in academia and in scholarly research. These endeavors, however, have infrequently offered a comprehensive examination of the breadth of research on SETs: types of bias, impact of inappropriate evaluation, interventions to reduce problems, and alternatives to evaluative systems that utilize SETs. This article uniquely addresses all of these aspects.

While there is an extensive literature on the problems with SETs and biases against certain instructors, there is no published comprehensive review of this literature. Instead, published work tends to cite only the literature that is most pertinent to the author's research question or study design. To construct this novel review, we draw on an original database of more than 90 articles on evaluative bias constructed from across academic disciplines[2]. Our dataset draws on scholarship ranging from 1974 to 2020 and relies on diverse methodological approaches, including experiments, SETs, Rate My Professor, other surveys, and experimental comparisons of sample syllabi.

This uniquely exhaustive approach to examining prior research allows us to offer a comprehensive look at the literature on bias in teaching evaluations to summarize and contextualize major findings on the two predominant types of bias: measurement bias and equity bias. We also identify gaps in understanding and suggest future research in hopes of offering insights that will assist faculty in advocating for and implementing institutional change. We highlight the understudied role of intersectionality, point to promising efforts to reduce bias, and offer recommendations for reform.

Overview of Major Findings

We would be remiss if we did not first acknowledge that there exists a disparate literature that includes some studies that have found no evidence of bias in gender evaluations (Basow & Distenfeld, 1985; Basow & Howe, 1987; Bennett, 1982; Elmore & LaPointe, 1974; Harris, 1975; Kaschak, 1981) or at least when controlling for other sources of bias (Centra & Gaubatz, 1998; Feldman, 1992; Theall & Franklin, 2001; Benton & Cashin, 2012; Wright & Jenkins-Guarnieri, 2012). There is evidence from the existing literature that the effect of gender on SETs is conditional. Some studies have, in fact, found women to be advantaged in evaluations, especially in departments where women are over-represented, such as certain humanities fields (Wigington et al., 1989; Rowden & Carlson, 1996; Whitworth et al., 2002). Other identities have been studied less frequently but there is some evidence an "ethnicity affinity" effect where students who share their instructor's ethnicity are more likely to rate them more highly.

There are a few meta-analyses of teaching evaluations, some of which consider potential gender biases. These analyses find mixed evidence of gender bias. For example, Wright and Jenkins-Guarnieri (2012) find that SETs are "largely free from gender bias." We believe these meta-analyses do not adequately assess the effect of gender. Meta-analyses

[2] A full list of articles and article summaries are available at < redacted >

require the comparison of similar or identical variables and models across studies. As a result, three types of evidence are generally excluded in these analyses: 1) quantitative data from sources such as RateMyProfessor.com, non-SET surveys, or questions from SETs that have unique question wording etc., 2) experimental designs that compare sample syllabi or online class experiences rather than SETs, 3) qualitative, open-answer sections of SETs, where sexist, racist or homophobic comments emerge.

We do not deny there are conditions in which those with marginalized identities might fare as well or even better in assessments by their students. It is the nuance around this multitude of competing variables we wish to add to the discussion of SETs. Furthermore, we acknowledge SETs may fulfil some of their intended purpose despite the potential for discrimination. At least one scholar (Murray, 1997) has demonstrated evaluations as being associated with improved teaching and likely don't contribute to other negative consequences like grade inflation.

In our analysis, two problematic and consistent findings predominate the literature. First, we find that scholars across disciplines and in numerous country contexts consistently reveal that SETs do not measure teaching effectiveness (Uttl et al., 2017; Benton & Cashin, 2012). That is, SETs are prone to measurement bias. Second, most of the literature indicates that men receive higher evaluative scores compared to women (see for instance, Basow & Silberg, 1987; MacNell et al., 2015; Mengel et al., 2018; Sidanius & Crane, 1989; Wigington et al., 1989). There is some evidence of discrimination towards other groups, as well, though it is less-well documented in the scholarship (as we will discuss below). In other words, SETs are also prone to equity bias. We review these two forms of bias below.

Measurement Bias. When variables unrelated to teaching effectiveness systematically influence the results in SETs, this constitutes an issue with measurement (Marsh, 1984; Benton & Cashin, 2012). Examples of variables that contribute to measurement bias include both course characteristics (such as class time, class size, if the class is a required or elective class, course difficulty, and discipline) and individual characteristics of students (like their interest in the class material and student's previous coursework). Despite the ubiquity of SETs across academia, research shows myriad ways in which SETs are poor indicators of teaching or course quality. Teaching evaluations are only weakly correlated or entirely uncorrelated with teaching effectiveness (Stark & Freishtat, 2014; Uttl et al., 2017). In their meta-study of evaluations, Utll et al. (2017) conclude, "…the best evidence–the meta-analyses of SET/learning correlations when prior learning/ability are taken into account–indicates that the SET/learning correlation is zero (19)." Using a simulation that assumes course evaluations are moderately correlated with learning (0.4), SETs identify the wrong instructor as the superior teacher 37% of the time (Esarey & Valdes, 2020).

Instead, instructors receive higher evaluations for classes with lighter workloads or higher grading distributions (Greenwald & Gillmore, 1997; Miles & House, 2015; Rosen, 2018; Sinclair & Kunda, 2000). Teaching evaluations are also higher when students are more engaged or excited by the teaching modalities and lower for non-elective and quantitative courses (Benton & Cashin, 2012; Boring et al., 2016; Chamberlin & Hickey, 2001; Elmore & LaPointe, 1975; Greenwald & Gillmore, 1997; Mengel et al., 2018; Uttl et al., 2017, 2013). Students give higher evaluations to upper-level, discussion-based classes over larger, introductory courses (Hamermesh & Parker, 2005; Miles & House, 2015; Sidanius & Crane, 1989; Spooren et al., 2013; Centra & Gaubatz, 1998). Furthermore, there are disciplinary differences across evaluations; natural science courses receive the lowest scores and humanities the highest (Basow & Montgomery,

2005; Basow & Silberg, 1987). Even bringing cookies or chocolate to class shapes course evaluations (Hessler et al., 2018; Youmans & Jee, 2007).

Essentially, evaluations are shaped by discipline, student interest, class level, class difficulty, class meeting time, and other course-specific characteristics, but not generally actual instructor quality (Franklin & Theall, 1995; Greenwald & Gillmore, 1997; Miles & House, 2015; Spooren et al., 2013; Wigington et al., 1989; Uttl et al., 2017, 2013; Wachtel, 1998). Computational simulation models demonstrate that "even under ideal conditions, under ideal circumstances, even careful and judicious use of SETs to assess faculty can produce an unacceptably high error rate" (Esarey & Valdes, 2020, 1). On this basis alone, universities and departments should reconsider how these evaluations are used in high stakes employment decisions, such as hiring and promotion for all instructors. These concerns are compounded for faculty who may face additional penalties because of their personal characteristics.

Equity Bias. When variables outside the instructor's control systematically influence the results, this is equity bias. Examples of equity bias include bias relating to the instructor's gender, race, ethnicity, accent, sexual orientation, or disability status. This is clearest in the literature demonstrating instructor gender influences evaluative outcome, especially in qualitative comments about the course or instructor. While a few meta-studies find few gender differences (Wallisch & Cachia, 2019; Wright & Jenkins-Guarnieri, 2012), the vast body of literature across time and methodological approach consistently finds the opposite. Research has demonstrated a multitude of ways that men benefit from evaluation, while women do not fare as positively. For example, men are more likely to be seen as adept, brilliant, or organized instructors across a range of institutional settings and research methodologies (Arbuckle & Williams, 2003; Abel & Meltzer, 2007; Boring et al., 2016; MacNell et al., 2015; McPherson et al., 2009; Mengel et al., 2018; Ridgeway, 2011; Sidanius & Crane, 1989; Wagner et al., 2016).

There is, however, evidence that these differences are less pronounced in some disciplines than in others. For instance, Basow and Montgomery (2005) find that women receive lower scores in the social sciences, but women's scores are higher in the humanities. However, Rosen (2018), using a massive ($n = 7{,}800{,}000$) Rate My Professor sample, finds there is no discipline where women receive higher evaluative scores. These lower scores may also be substantial, with one study finding that, controlling for other factors, female instructors received average ratings that were one-half standard deviation lower than men's ratings (Hamermesh & Parker, 2005).

Specific conditions where women are at a disadvantage over their male colleagues abound. Disparate research demonstrates that men are perceived as more accurate in their teaching, have higher levels of education, are less sexist, more enthusiastic, competent, organized, professional, effective, easier to understand, prompt in providing feedback, and are less-harshly penalized for being tough graders (Abel & Meltzer, 2007; Arbuckle & Williams, 2003; Basow, 1995; Basow & Silberg, 1987; Boring et al., 2016; Elmore & LaPointe, 1975; MacNell et al., 2015; Rivera & Tilcsik, 2019; Miller & Chamberlin, 2000; Sidanius & Crane, 1989; Sinclair & Kunda, 2000; Sprague & Massoni, 2005; Storage et al., 2016). Experimental designs that manipulate the gender of the instructor in online teaching environments have even shown that students offered lower evaluations when they believed the instructor was a woman, despite identical course delivery (Boring et al., 2016; MacNell et al., 2015). Students are also more likely to expect special favors from female professors and react badly when those expectations aren't met or fail to follow directions when they are offered by a woman professor (El-Alayli et al., 2018; Piatak & Mohr, 2019).

A powerful consideration is that women and men appear to be evaluated—by students of both genders—through the lens of gender stereotypes (Wallisch & Cachia, 2019). Women are rated highly for exhibiting traditionally-feminine traits (Eagly & Karau, 2002) like warmth and sensitivity (Bennett, 1982; Kierstead et al., 1988; Sidanius & Crane, 1989), while men are evaluated positively on gendered perceptions of their intellectual and teaching prowess (Bian et al., 2017; Basow, 2000; Boring, 2017; Leslie et al., 2015). Students prefer professors to have masculine traits (Burns-Glover & Veith, 1995) and penalize women professors when they fail to conform to gender stereotypes (Bennett, 1982; Boring, 2017; Kierstead et al., 1988). In fact, conforming to prescribed gender roles has a more pronounced effect on evaluations than gender itself does (Basow and Silberg, 1987; Freeman, 1994).

Also relevant to how male and female instructors are evaluated is the gender of the student conducting the evaluation. Not unsurprisingly, research finds evidence of a gender-affinity effect, with students rating faculty that share their gender more highly (Bachen et al., 1999; Martin, 1984; Young et al., 2009). Male students rate their female instructors lower (Basow & Silberg, 1987; Basow, 1995; Burns-Glover & Veith, 1995; Fan et al., 2019; Kaschak, 1981, 1978; Mengel et al., 2018) while female students rate them higher (Bray & Howard, 1980; Centra, 2000; Rowden & Carlson, 1996)[3]. Abel and Meltzer (2007) also demonstrate that a student's sexism predicts how they will evaluate women and how they will perceive instructor sexism, though this finding was only significant for conservative students[4].

While the evidence on gender discrimination is strong, we find there is substantially less research on the possibility of bias in teaching evaluations for faculty color, in no small part because of their severe underrepresentation in academia (APSA, 2011). However, the research that does exist suggests that faculty of color are evaluated worse than their white colleagues (Reid, 2010), especially Black and Asian professors, with Black men faring particularly poorly. Additionally, faculty with accents fare worse than their white and native English-speaking counterparts. For instance, Smith and Hawkins (2011) show that Black and other non-White faculty received the lowest mean scores across 26 individual multidimensional evaluation items as well as two global measures of course quality, overall value, and overall teaching ability (see also Hamermesh & Parker, 2005). Faculty with accents and Asian last names receive lower ratings in both SETs and Rate My Professor (Fan et al., 2019; Subtirelu, 2015). People of color may also be punished more for intersectional stereotype nonconformity (Anderson, 2010); Latina women are perceived less warmly than Anglo women with similarly strict teaching style (Anderson & Smith, 2005) and women of color are evaluated more harshly than white men (Chávez & Mitchell, 2020).

There is also scant research on biases towards LGBT faculty; however, experimental evidence suggests they may be perceived as more politically biased than heterosexual faculty (Anderson & Kanner, 2011). In an experiment, students were more likely to rate gay and lesbian instructors who were strong lecturers lower than lecturers with an unspecified orientation,

[3] Though see Basow and Montgomery (2005), which finds no significant interactions between student and faculty gender

[4] Research also finds that the role of attractiveness is more relevant to women, who are more likely to get comments about their appearance (Mitchell & Martin, 2018; Key & Ardoin, 2019). This is problematic given that attractiveness has been shown to be correlated with evaluations of instructional quality (Rosen, 2018)

though they also rated gays and lesbians who were weak lecturers more moderately (Ewing et al., 2003).

It is even less clear how biases affect faculty across rank and age. Some research indicates that seniority decreases bias (Mengel et al., 2018; Wigington et al., 1989), while other research finds that younger professors are more popular and receive higher evaluations (Arbuckle & Williams, 2003; McPherson et al., 2009). We know almost nothing about biases against other relevant intersectional identities, such as disability or pregnancy and motherhood (but see Baker & Copp, 1997).

Recommendations for Better Evaluation

There are many well-documented ways that measurement bias and equity bias shape student evaluations of teaching. Given these biases and their role in personnel decisions, many colleges and universities – sometimes at the behest of faculty unions – are reevaluating how these ratings are used. We argue that we need not "throw out the baby with the bathwater" and eliminate the role of student evaluations entirely. Rather, they should be properly contextualized and used with caution. We offer the following six actions that individual faculty members and universities can take to make the use of student ratings more responsible.

1. Contextualize evaluations as perceptions of student learning, not as a measure of actual teaching.

   Although they are commonly called "student evaluations of teaching," SETs do not actually evaluate teaching. Instead, student evaluations represent their perception or experiences in a course (Linse, 2017; Abrami, 2001; Arreola, 2004). Students should not, and arguably cannot, evaluate teaching. A more accurate name for these experiences would be student experience questionnaires or student perceptions of learning. When properly contextualized as feedback on experience, rather than evaluating teaching, these assessments can provide useful feedback for faculty and administrators.
2. Be proactive about increasing the validity of the assessment by improving response rates.

   Administrators should not distribute or use assessments based on a low response rate. A low response rate makes it more likely that the sample is unrepresentative, which calls into question the validity of the assessment (Chapman & Joines, 2017; Adams & Umbach, 2012). Faculty can improve the response rate on evaluations by providing time for students to complete them in class, even if the evaluations are distributed online. Faculty should leave the room while this takes place, to alleviate pressure on students. Faculty can also improve the response rate by discussing the purpose of evaluations and how they can lead to improvements in the course for future students (Linse, 2017; Chapman & Joines, 2017).
3. Administrators should interpret the results of student ratings with caution.

   Student evaluations are not designed to be used as a comparative metric across faculty (Franklin, 2001); rather, their purpose is to gather information about how students perceived a faculty member teaching a certain course. As such, evaluations should be used to compare a faculty member's trajectory of teaching over time, and ideally, within a single course (Linse, 2017). Because one way that equity bias manifests is through lower evaluations for astereotypic instructors (i.e., women in male-dominated fields and vice versa), comparisons across faculty members further disadvantage already marginalized faculty.

Most faculty members receive mostly positive reviews (Linse, 2017). In fact, most faculty members' reviews do not follow a normal distribution and have a negative skew (Arreola, 2004; Ḥaṭiva, 2013a, b). The mean of a skewed distribution is more influenced by outliers, especially in smaller samples. In student ratings, outliers are usually negative (Linse, 2017). Administrators should look at the overall distribution of ratings rather than the mean, or look at the median or modal response rather than the mean. Furthermore, they should focus on trends and patterns rather than focusing on the individual numbers. It's common for ratings to vary as much as 0.4 points, depending on the scale (Linse, 2017; Marsh, 1980, 1982a, b).

Finally, administrators should report several ratings from several questions on the assessment rather than relying on a single global question about overall teaching effectiveness to reduce the effect of measurement error (Fischer & Hänze, 2019; Smith & Hawkins, 2011).

4. Restrict or eliminate the use of qualitative comments.

   Across all the studies in our sample, the clearest evidence of gender bias is in qualitative comments. Scholars employing content analysis of qualitative comments finds clear evidence of bias with women faculty and faculty of color are more likely to receive negative comments about personality traits, appearance, mannerisms, competence, and professionalism compared to white men (Wallace et al., 2019). Furthermore, many faculty report particularly mean-spirited and cruel comments (Lindahl & Unger, 2010). Instead of asking for general "comments," assessments should direct students to provide feedback on certain experiences with the course, as this may reduce irrelevant and mean comments.

   There are additional problems with qualitative comments beyond issues of bias. They are difficult to aggregate and have a low sample size (Himelein, 2018). Furthermore, they are not reliable—in fact, they frequently have contradictory feedback (Linse, 2017). Finally, even well-intentioned reviewers of qualitative comments may be susceptible to novelty bias (we are more likely to remember unexpected or uncommon findings) and negativity bias (the tendency to be influenced by negative information more than positive information) (Himelein, 2018). Comments that are anomalous or do not correlate with class averages on quantitative items should be disregarded.

5. Administrators must not rely on student evaluations as the sole method of assessing teaching.

   There are many ways to evaluate faculty teaching beyond standard student evaluations. A few of the most common alternatives to standard evaluations of teaching include peer observations (Miller & Seldin, 2014), comprehensive evaluations of teaching portfolios (Centra, 2000; Seldin et al., 2010), and internal or external reviews of course materials (Chism, 2007). While it's worth noting that these alternative methods may also be prone to some of the same biases, using multiple (potentially) flawed measures of teaching is better than a single measure, provided they aren't all systematically biased in the same way (Esarey & Valdes, 2020). It's true that these alternatives are more laborious than SETs, but we ought not to rely on a problematic measure simply because it is easier.

6. Produce more research in interventions to reduce bias.

   While there are multiple dozens of articles establishing bias in student evaluations of teaching, there are very few articles that test interventions to mitigate bias. What little research exists yields some promising leads. For instance, reducing the size of the scale can mitigate gender bias (Rivera & Tilcsik, 2019). Another study that uses a randomized control trial finds that making students aware of biases can mitigate the gender gap in SETs (Peterson et al., 2019), though the evidence here is somewhat contradictory (Key

& Ardoin, 2019), and anecdotally induce a backlash effect. Administrators and faculty alike would also benefit from more specific insights into measures that provoke less bias, approaches that increase sample size on evaluations, appropriate means for triangulating assessment methods, enhanced understanding of how to interpret results, and many other facets of evaluation that have received little or no scholarly attention. Clearly, this is an area ripe for future research.

Conclusion

It is clear that teaching evaluations are poor metrics of student learning and are, at best, imperfect measures of instructor performance. SETs disproportionately penalize faculty who are already marginalized by their status as minority members of the discipline. Across the existing literature, using different data, measures, and methods, scholars in many disciplines have documented problems with student evaluations of teaching in ways that are abundantly relevant to faculty in all disciplines.

There are steps that faculty and administrators can take to reduce measurement and equity bias in evaluations of teaching and the pernicious use of student evaluations in critical personnel decisions. While the literature testing interventions and strategies to mitigate biases is relatively nascent, it is promising. More research should be done to rigorously test interventions that improve the quality and fairness of assessments. Until feasible, reliable, and fair methods for evaluating teaching and learning are established, more caution should be taken in the use of SETs in hiring, tenure, and promotion decisions and alternatives assessments of teaching should be further utilized.

Declarations

Conflict of Interest The authors hereby acknowledge no financial or non-financial conflict of interest.

References

Abel, M. H., & Meltzer, A. L. (2007). Student ratings of a male and female professors' lecture on sex discrimination in the workforce. *Sex Roles, 57*(3–4), 173–180

Abrami, P. C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. *New Directions for Institutional Research, 2001*(109), 59–87

Adams, M. J. D., & Umbach, P. D. (2012). Nonresponse and online student evaluations of teaching: understanding the influence of salience, fatigue, and academic environments. *Research in Higher Education, 53*(5), 576–591

Anderson, K. J. (2010). Students' stereotypes of professors: An exploration of the double violations of ethnicity and gender. *Social Psychology of Education, 13*(4), 459–472

Anderson, K. J., & Kanner, M. (2011). Inventing a Gay Agenda: Students' Perceptions of Lesbian and Gay Professors 1. *Journal of Applied Social Psychology, 41*(6), 1538–1564

Anderson, K. J., & Smith, G. (2005). Students' preconceptions of professors: Benefits and barriers according to ethnicity and gender. *Hispanic Journal of Behavioral Sciences, 27*(2), 184–201

Aguirre Jr, A. (2000). *Women and Minority Faculty in the Academic Workplace: Recruitment, Retention, and Academic Culture. ASHE-ERIC Higher Education Report, Volume 27, Number 6. Jossey-Bass Higher and Adult Education Series.* Jossey-Bass, 350 Sansome St., San Francisco, CA 94104-1342.

APSA. (2011). Political science in the 21st century edited by report of the task force on political science in the 21st century

Arbuckle, J., & Williams, B. D. (2003). Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. *Sex Roles, 49*(9–10), 507–516

Arreola, R. A. (2004). *Developing a comprehensive faculty evaluation system.* Magna Publications
Bachen, C. M., McLoughlin, M. M., & Garcia, S. S. (1999). Assessing the role of gender in college students' evaluations of faculty. *Communication Education, 48*(3), 193–210
Baker, P., & Copp, M. (1997). Gender matters most: the interaction of gendered expectations, feminist course content, and pregnancy in student course evaluations. *Teaching Sociology*: 29–43
Barbezat, D. A., & Hughes, J. W. (2005). Salary structure effects and the gender pay gap in academia. *Research in Higher Education, 46*(6), 621–640.
Bos, A. L., Sweet-Cushman, J., & Schneider, M. C. (2019). Family-friendly academic conferences: a missing link to fix the "leaky pipeline"? *Politics, Groups, and Identities, 7*(3), 748–758.
Basow, S. A., & Distenfeld, M. S. (1985). Teacher expressiveness: More important for male teachers than female teachers? *Journal of Educational Psychology, 77*(1), 45
Basow, S. A., & Howe, K. G. (1987). Evaluations of college professors: Effects of professors' sex-type, and sex, and students' sex. *Psychological Reports, 60*(2), 671–678
Basow, S. A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology, 87*(4), 656
Basow, S. A. (2000). Best and worst professors: Gender patterns in students' choices. *Sex Roles, 43*(5–6), 407–417
Basow, S. A., & Montgomery, S. (2005). Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education, 18*(2), 91–106
Basow, S. A., & Silberg, N. T. (1987). Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology, 79*(3), 308
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*(2), 170
Benton, S. L., & Cashin, W. E. (2012). Student ratings of teaching: a summary of research and literature (IDEA Paper no. 50). Manhattan, KS: The IDEA Center
Bian, L., Leslie, S.-J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science, 355*(6323), 389–391
Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics, 145,* 27–41
Boring, A., Ottoboni, K., & Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*
Bray, J. H., & Howard, G. S. (1980). Interaction of teacher and student sex and sex role orientations and student evaluations of college instruction. *Contemporary Educational Psychology,5*(3), 241–248
Burns-Glover, A. L., & Veith, D. J. (1995). Revisiting gender and teaching evaluations: Sex still makes a difference. *Journal of Social Behavior and Personality, 10*(4), 69
Centra, J. A. (2000). Evaluating the Teaching Portfolio: A Role for Colleagues. *New Directions for Teaching and Learning, 83,* 87–93
Centra, J. A., & Gaubatz, N. B. (1998). Is there gender bias in student ratings of instruction. *Journal of Higher Education, 70,* 17–33
Chamberlin, M. S., & Hickey, J. S. (2001). Student evaluations of faculty performance: The role of gender expectationis in differential evaluations. *Educational Research Quarterly, 25*(2), 3
Chapman, D. D., & Joines, J. A. (2017). Strategies for Increasing Response Rates for Online End-of-Course Evaluations. *International Journal of Teaching and Learning in Higher Education, 29*(1), 47–60
Chávez, K., & Mitchell, K. M. (2020). Exploring bias in student evaluations: Gender, race, and ethnicity. *PS: Political Science & Politics*, *53*(2), 270-274.
Chism, N. V. N. (2007). *Peer Review of Teaching. A Sourcebook*. Bolton Massachusetts: Anker
Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review, 109*(3), 573
El-Alayli, A., Hansen-Brown, A. A., & Ceynar, M. (2018). Dancing backwards in high heels: Female professors experience more work demands and special favor requests, particularly from academically entitled students. *Sex Roles, 79*(3–4), 136–150
Elmore, P. B., & LaPointe, K. A. (1974). Effects of teacher sex and student sex on the evaluation of college instructors. *Journal of Educational Psychology, 66*(3), 386.
Elmore, P. B., & LaPointe, K. A. (1975). Effect of teacher sex, student sex, and teacher warmth on the evaluation of college instructors. *Journal of Educational Psychology, 67*(3), 368
Esarey, J., & Valdes, N. (2020). Unbiased, reliable, and valid student evaluations can still be unfair. *Assessment & Evaluation in Higher Education*

Ewing, V. L., Stukas Jr, A. A., & Sheehan, E. P. (2003). Student prejudice against gay male and lesbian lecturers. *The Journal of Social Psychology, 143*(5), 569–579
Fan, Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., & Johnston, E. L. (2019). Gender and cultural bias in student evaluations: Why representation matters. *PLoS One, 14*(2), e0209749
Feldman, K. 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
Fischer, E., & Hänze, M. (2019). Bias hypotheses under scrutiny: investigating the validity of student assessment of university teaching by means of external observer ratings. *Assessment & Evaluation in Higher Education, 44*(5), 772–786
Franklin, J. (2001). Interpreting the numbers: Using a narrative to help others read student evaluations of your teaching accurately. *New Directions for Teaching and Learning, 87,* 85–100
Franklin, J., & Theall, M. (1995). The relationship of disciplinary differences and the value of class preparation time to student ratings of teaching. *New Directions for Teaching and Learning, 1995*(64), 41–48
Freeman, H. R. (1994). Student evaluations of college instructors: Effects of type of course taught, instructor gender and gender role, and student gender. *Journal of Educational Psychology, 86*(4), 627
Greenwald, A. G., & Gillmore, G. M. (1997). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89*(4), 743
Hamermesh, D. S., & Parker, A. (2005). Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review, 24*(4), 369–376
Harris, M. B. (1975). Sex role stereotypes and teacher evaluations. *Journal of Educational Psychology, 67*(6), 751
Ḥaṭiva, N. (2013a). *Student ratings of instruction: a practical approach to designing, operating, and reporting*. Oron Publications
Ḥaṭiva, N. (2013b). *Student ratings of instruction: Recognizing effective teaching*. Oron Publications
Hessler, M., Pöpping, D. M., Hollstein, H., Ohlenburg, H., Arnemann, P. H., Massoth, C., et al. (2018). Availability of cookies during an academic course session affects evaluation of teaching. *Medical Education, 52*(10), 1064–1072
Himelein, M. J. (2018). Pitfalls of using student comments in the evaluation of faculty. *Academic Briefing: Expert Advice for Higher Ed Leaders.* https://www.academicbriefing.com/human-resources/faculty-evaluation/pitfalls-of-using-student-comments-evaluation-of-faculty/
Kaschak, E. (1978). Sex bias in student evaluations of college professors. *Psychology of Women Quarterly, 2*(3), 235–243
Kaschak, E. (1981). Another look at sex bias in students' evaluations of professors: Do winners get the recognition that they have been given? *Psychology of Women Quarterly, 5*(5_suppl), 767–772
Key, E., & Ardoin, P. (2019). Students rate male instructors more highly than female instructors. We tried to counter that hidden bias. *Washington Post*. Accessed 3 Sep 2019. https://www.washingtonpost.com/politics/2019/08/20/students-rate-male-instructors-more-highly-than-female-instructors-we-tried-counter-that-hidden-bias/
Kierstead, D., D'agostino, P., & Dill, H. (1988). Sex role stereotyping of college professors: Bias in students' ratings of instructors. *Journal of Educational Psychology, 80*(3), 342
Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science, 347*(6219), 262–265
Lindahl, M. W., & Unger, M. L. (2010). Cruelty in student teaching evaluations. *College Teaching, 58*(3), 71–76
Linse, A. 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
MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education, 40*(4), 291–303
Marsh, H. W. (1980). Research on students' evaluations of teaching effectiveness. *Instructional Evaluation, 4*(5), 5–13
Marsh, H. W. (1982a). Factors affecting students' evaluations of the same course taught by the same instructor on different occasions. *American Educational Research Journal, 19*(4), 485–497
Marsh, H. W. (1982b). Validity of students' evaluations of college teaching: A multitrait–multimethod analysis. *Journal of Educational Psychology, 74*(2), 264
Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential baises, and utility. *Journal of Educational Psychology, 76*(5), 707
Martin, E. (1984). Power and authority in the classroom: Sexist stereotypes in teaching evaluations. *Signs: Journal of Women in Culture and Society, 9*(3), 482–492
McPherson, M. A., Todd Jewell, R., & Kim, M. (2009). What determines student evaluation scores? A random effects analysis of undergraduate economics classes. *Eastern Economic Journal, 35*(1), 37–51

Mengel, F., Sauermann, J., & Zölitz, U. (2018). Gender bias in teaching evaluations. *Journal of the European Economic Association, 17*(2), 535–566

Miles, P., & House, D. (2015). The Tail Wagging the Dog; An Overdue Examination of Student Teaching Evaluations. *International Journal of Higher Education, 4*(2), 116–126

Miller, J., & Seldin, P. (2014). Changing Practices in Faculty Evaluations: Can Better Evaluation Make a Difference? *Academe, 100*(3), 35–38

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

Mitchell, K. M. W., & Martin, J. (2018). Gender bias in student evaluations. *Political Science & Politics, 51*(3), 648–652

Murray, H. G. (1984). The impact of formative and summative evaluation of teaching in North American universities. *Assessment and Evaluation in Higher Education, 9*(2), 117–132

Murray, H. G. (1997). Does evaluation of teaching lead to improvement of teaching? *The International Journal for Academic Development, 2*(1), 8–23.

Perna, L. W. (2005). The benefits of higher education: Sex, racial/ethnic, and socioeconomic group differences. *The Review of Higher Education, 29*(1), 23–52.

Peterson, D. A. M., Biederman, L. A., Andersen, D., Ditonto, T. M., & Roe, K. (2019). Mitigating gender bias in student evaluations of teaching. *PLoS One, 14*(5), e0216241

Piatak, J., & Mohr, Z. (2019). More gender bias in academia? Examining the influence of gender and formalization on student worker rule following. *Journal of Behavioral Public Administration, 2*(2)

Reid, L. D. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors. Com. *Journal of Diversity in Higher Education, 3*(3), 137

Ridgeway, C. L. (2011). *Framed by gender: How gender inequality persists in the modern world* Oxford University Press

Rivera, L. A., & Tilcsik, A. (2019). Scaling Down Inequality: Rating Scales, Gender Bias, and the Architecture of Evaluation. *American Sociological Review, 84*(2), 248–274

Rosen, A. 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

Rowden, G. V., & Carlson, R. E. (1996). Gender issues and students' perceptions of instructors' immediacy and evaluation of teaching and course. *Psychological Reports, 78*(3), 835–839

Seldin, P., Miller, J. E., & Seldin, C. A. (2010). *The teaching portfolio: A practical guide to improved performance and promotion/tenure decisions*. John Wiley & Sons

Sidanius, J., & Crane, M. (1989). Job evaluation and gender: The case of university faculty. *Journal of Applied Social Psychology, 19*(2), 174–197

Sinclair, L., & Kunda, Z. (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.

Smith, B. P., & Hawkins, B. (2011). Examining student evaluations of black college faculty: does race matter? *Journal of Negro Education, 80*(2)

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research, 83*(4), 598–642

Sprague, J., & Massoni, K. (2005). Student evaluations and gendered expectations: What we can't count can hurt us. *Sex Roles, 53*(11–12), 779–793

Stark, P., & Freishtat, R. (2014). An evaluation of course evaluations. ScienceOpen. *Center for Teaching and Learning, University of California, Berkley.* Retrieved https://www.scienceopen.com/document

Storage, D., Horne, Z., Cimpian, A., & Leslie, S.-J. (2016). The frequency of "brilliant" and "genius" in teaching evaluations predicts the representation of women and African Americans across fields. *PLoS One, 11*(3), e0150194

Subtirelu, N. C. (2015). "She does have an accent but…": Race and language ideology in students' evaluations of mathematics instructors on RateMyProfessors. com. *Language in Society, 44*(1), 35–62

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? *New Directions for Institutional Research, 2001*(109), 45–56

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54,* 22–42

Uttl, B., White, C. A., & Morin, A. (2013). The numbers tell it all: students don't like numbers! *PLoS One, 8*(12), e83443

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 23*(2), 191–212

Wagner, N., Rieger, M., & Voorvelt, K. (2016). Gender, ethnicity and teaching evaluations: Evidence from mixed teaching teams. *Economics of Education Review, 54,* 79–94

Wallace, S. L., Lewis, A. K., & Allen, M. D. (2019). The State of the Literature on Student Evaluations of Teaching and an Exploratory Analysis of Written Comments: Who Benefits Most? *College Teaching, 67*(1), 1–14

Wallisch, P., & Cachia, J. (2019). Determinants of perceived teaching quality: the role of divergent interpretations of expectations

Wigington, H., Tollefson, N., & Rodriguez, E. (1989). Students' ratings of instructors revisited: Interactions among class and instructor variables. *Research in Higher Education, 30*(3), 331–344

Whitworth, J. E., Price, B. A., & Randall, C. H. (2002). Factors that affect college of business student opinion of teaching and learning. *Journal of Education for Business, 77*(5), 282–289

Wright, S. L., & Jenkins-Guarnieri, M. A. (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

Youmans, R. J., & Jee, B. D. (2007). Fudging the numbers: Distributing chocolate influences student evaluations of an undergraduate course. *Teaching of Psychology, 34*(4), 245–247

Young, S., Rush, L., & Shaw, D. (2009). Evaluating Gender Bias in Ratings of University Instructors' Teaching Effectiveness. *International Journal for the Scholarship of Teaching and Learning, 3*(2), n2