2018

Annotated literature review: student evaluations of teaching (SET)

Eva, Nicole

University of Lethbridge Faculty Association

UNIVERSITY OF LETHBRIDGE FACULTY ASSOCIATION

Room D620, University Hall
4401 University Drive West, Lethbridge, AB T1K 3M4
www.ulfa.ca
Phone: 403 329 2578 Fax: 403 329 2113

Annotated Literature Review – Student Evaluations of Teaching (SET)
Prepared by ULFA's Gender, Equity & Diversity Committee, 2016-2018.*

ULFA's statement on the Use of Student Evaluations of Teaching may be found here:
http://ulfa.ca/statement-on-student-evaluations-on-teaching/

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. https://doi.org/10.1177/0739986304273707

This study found that gender, ethnicity, and teaching style interact in ways such that instructors are evaluated more highly when their style corresponds with student expectations, and that Latino women were penalized more heavily than White women for not displaying warmth. Anglo women also received higher ratings for capability, despite having the same teaching style as Latinas.

Arbuckle, J., & Williams, B. D. (2003). Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. *Sex Roles: A Journal of Research*, *49*(9/10), 507–516. https://doi.org/10.1023/A:1025832707002

This paper uses a natural experiment to examine the relationship between college students' perceptions of professors' expressiveness and implicit age and gender stereotypes. The experiment consists of three hundred and fifty-two male (154) and female (198) students that watch a thirty-five-minute picture slide-audio taped presentation of a computer-generated genderless stick figure. The audiotape was read by a 45-year old woman whose voice was neutral, and did not reveal her gender and age. The stick figure and neutral voice allowed the authors to control for physical appearance, teaching style, gender, age, and tone of voice. After the lecture, the students were told the instructor was either a young female, a young male, an older female, or an older male. Students who thought the instructor was a young male rated the lecture highest, because they thought (he) was expressive and enthusiastic professors. This effect was seen by both male and female students.

Baldwin, T., & Blattner, N. (2003). Guarding against potential bias in student evaluations: What every faculty member needs to know. *College Teaching, 51*(1), 27–32. Retrieved from https://doi.org/10.1080/0260293980230207

This article is useful from an educational standpoint, in that it flags the issues that faculty and evaluators should know before placing too much emphasis on SET. It explains how student evaluations are suspect enough in their validity as to not be weighed heavily in faculty evaluation processes, and mentions that SETs were originally intended for faculty use in improving their own teaching, and that they are being misused. Other solutions are proposed.

UNIVERSITY OF LETHBRIDGE FACULTY ASSOCIATION

Room D620, University Hall
4401 University Drive West, Lethbridge, AB T1K 3M4
www.ulfa.ca
Phone: 403 329 2578 Fax: 403 329 2113

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), 308-314. https://doi.org/10.1037/0022-0663.79.3.308

This experiment looked at the evaluations by 1000 male and female college students of 16 male and female professors. They evaluated their instructors in terms of teaching effectiveness and sex-typed characteristics. Male students gave female professors significantly poorer ratings than they gave male professors; their ratings of female professors were poorer than those of female students on four of the six measures. Female students also evaluated female professors less favourably than male professors on three measures. Student perceptions of a professor's instrumental/active and expressive/nurturant traits, which were positively related to student ratings of teaching, accounted for only a few of these gender-related effects. Student major and student class standing also played a role in the evaluation of professors.

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

The author looked at evaluations completed over a 4-year period at a private liberal arts college. They were analysed for the effects of teacher gender, student gender, and divisional affiliation (humanities, social science, natural science, engineering, interdisciplinary). They found that ratings of male professors appeared to be unaffected by student gender; female professors tended to receive their highest ratings from female students and their lowest ratings from male students, even taking into consideration rank. Mean ratings received by female and male professors also varied as a function of the divisional affiliation. Female faculty tend to be rated highly by their female students, especially in the humanities, but less positively by their male students, especially in the social sciences. Male faculty were almost always rated higher on questions of knowledge; female faculty were generally rated higher on questions of respect, sensitivity, and student freedom to express ideas, especially by female students. Female teachers tend to be rated relatively low by male students on thought stimulation, appropriate speech, fairness, non repetition, and their overall rating. The author concludes that gender effects on student evaluations of professors in a naturalistic setting are complex. It is important to examine teacher gender along with such factors as student gender, the discipline of the course, and the specific questions on the form. Instructors whose particularities all are correlated negatively with student ratings may be particularly affected (e.g., gender, gender of students, teaching style, gender-typed personality characteristics, and discipline).

Basow, S. A. (2000). Best and worst professors: Gender patterns in students' choices. *Sex Roles*, *43*(5/6), 407–417. https://doi.org/10.1023/A:1026655528055

This article has students describe the qualities of their 'best' and 'worst' professor. Female students were more likely to rate female professors as their 'best'; male students chose females in smaller proportion than would be expected. For 'worst' professor, both males and females chose professors of either sex in expected proportions. In describing their 'best' professor,

Room D620, University Hall
4401 University Drive West, Lethbridge, AB T1K 3M4
www.ulfa.ca
Phone: 403 329 2578 Fax: 403 329 2113

students chose traits like dynamism/enthusiasm, and 'worst' professors did poorly on organization/clarity. It seems counterintuitive that clarity would not be a determining factor with a 'best' professor, yet it fell far short of the 'dynamism' element. This suggests that innate personality characteristics count more towards a professor's likeability than their actual teaching skills. Previous research shows interpersonal skills get higher evaluations, but in this study that was only true for female faculty. The traits that women stereotypically possess are typically not those that are rewarded by higher evaluations, and both women and men that go 'against type' are usually penalized. Contains a refutation to Feldman (micro vs macro). Limitation: small convenience sample.

Bates, L. (2015, February 13). Female academics face huge sexist bias – no wonder there are so few of them. *The Guardian*. Retrieved from https://www.theguardian.com/lifeandstyle/womens-blog/2015/feb/13/female-academics-huge-sexist-bias-students

Brief review of Benjamin Schmidt's research which plots language used in RateMyProfessors.com for women vs men. Women were more likely to be described as annoying, shrill, or strict while men were more likely to be described as smart, funny, or assertive. Different standards for each gender; desirable traits in one gender were not desirable in others. This supports Felton's (see below) 2008 study.

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), 170–179. http://dx.doi.org/10.1037/0022-0663.74.2.170

This paper uses Factor Analysis to analyze the students' perceptions as well as expectations for male and female instructors from nonscience introductory courses at a liberal arts college, using completed questionnaire from 253 students that included formal teaching performance ratings, perceptual orientation scales, and indicators of degree and context of student-instructor contact. A total of 11 female-instructed and 28 male-instructed courses were used in the analysis in which they were placed within a unitary frame of reference. The main four factors loading consider in this paper are: (1) Nonauthoritarian Interpersonal style; (2) Charisma (Potency); (3) Self-assurance ("Experienced"); and (4) Instructional Approach ("Professionalism"). The main findings of the analysis are: (1) There is no evidence of direct bias in formal student evaluation of instructors in this specific sample. (2) Evidence of gender differentiated standards is found in the area of expectations and judgments regarding willingness to assist and satisfaction with instructor's availability. That is, male instructors are judged independently of students' personal experiences of contact and access, whereas female instructors are judged far more closely. (3) Students demand a higher standard of formal preparation and organization from female instructors.

Berk, R. A. (2014). Should student outcomes be used to evaluate teaching? *Journal of Faculty Development, 28*(2), 87–96. Retrieved from https://www.google.ca/url?sa=t&rct=j&q=

Room D620, University Hall
4401 University Drive West, Lethbridge, AB T1K 3M4
www.ulfa.ca
Phone: 403 329 2578 Fax: 403 329 2113

&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjU7qKW7IzYAhUP6GMKHWKQA-sQFggqMAA&url=http%3A%2F%2Fwww.ronberk.com%2Farticles%2F2014_outcomes.pdf&usg=AOvVaw1vh0hDTc1LlQCTvmXIvssO

Berk observes that every source of evaluation of teaching is fallible. Instead of relying upon a single source, Berk argues in favour of picking the best sources from among the collection of available sources for the specific decision at hand. While each source may be fallible, in combination the strengths of each source can compensate for weaknesses of other sources. Berk argues that student ratings, despite their flaws, should be part of that mix of sources. Fifteen potential sources of evidence of teaching effectiveness are listed, but are described in greater detail in a separate papers (Berk 2006, 2013).
Much of the paper discusses the challenges in using learning outcomes as a measure of teaching effectiveness. These include identifying satisfactory measures of student achievement, or growth, and measuring the contribution of the instructor towards student achievement or growth. Berk analyses various types of outcome measures and their relationship to student ratings, finding little consensus as to the degree of association. Berk's related conclusion is that it is not currently possible to identify the teaching component of learning outcomes and gains. Berk ultimately argues that the use of learning outcomes are not appropriate in employment decisions about faculty, such as annual review and feedback, contract renewal, merit pay, promotion, and tenure.

Boring, A. (2017). Gender biases in student evaluations of teachers. *Journal of Public Economics, 145*(13), 27–41. https://doi.org/10.1016/j.jpubeco.2016.11.006

In this well-controlled experiment, the author analyzed a large database of evaluations from a French university. She found that female professors received lower SETs, that male students give higher scores to male professors, and that those ratings were higher than female students gave to either female or male professors. Despite this, students did equally well in the course, whether they were taught by males or females. Stereotyping was also found in that those who acted in accordance with the qualities expected by their gender were 'rewarded' with higher evaluations.

Boring, A., Dial, U. M. R., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, (January), 1–36. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

This paper uses two set of experiments of student evaluations of teaching (SET) to examine the teaching effectiveness and gender bias. The first experiment was conducted using SET data collected from 2004-2008 of first-year students in 1,177 sections a French university taught by 379 instructors (of which 34% are female). The second experiment was conducted using students in an online course at a U.S. university which are randomized into six sections of approximately 12 students in each section. For this particular experiment, two sections were taught by primary professors, two section were taught my male graduate teaching assistants

Room D620, University Hall
4401 University Drive West, Lethbridge, AB T1K 3M4
www.ulfa.ca
Phone: 403 329 2578 Fax: 403 329 2113

(TA), and the last two sections were taught by two female graduate TA. In one of the two sections taught by each TA, the TA uses his/her true name; whilst in the other she/he uses the other TA identity. Student has no direct contact with TAs, and the primary interactions were through online discussion boards.
The authors use nonparametric permutation stratified tests to test the hypothesis of the differences on the variety of means scores of SET between male and female instructors. Based on the two samples of data, their results suggest the following: (1) Among other things, SET are biased against female instructors, and the bias varies by discipline and by student gender. (2) The bias depends on many factors implying that it is not possible to correct for the bias. (3) SET are more sensitive to student's gender and grade expectation than they are to teaching effectiveness. (4) SET bias can cause more effective instructors to get lower SET than less effective instructors.

Boyd, S. (2003). Foreign-born teachers in the multilingual classroom in Sweden: The role of attitudes to foreign accent. *International Journal of Bilingual Education and Bilingualism, 6*(3-4), 283–295. https://doi.org/10.1080/13670050308667786

This paper looks at attitudes of teachers, students and administrators towards foreign-born teachers in the K-12 system. Specifically, it examined how negative attitudes created roadblocks to employment and advancement. The paper's focus is specific to the Swedish school system, examined during a period of considerable cultural change due to increasing immigration. It looked specifically at 4 subjects/languages: two Russian, one Hungarian and one Farsi. The study found that that the degree of accentedness played an important role in the listener's judgment of the teacher's language proficiency and suitability as a teacher

Boysen, G. A., Kelly, T. J., Raesly, H. N., & Casner, R. W. (2014). The (mis)interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education*, *39*(6), 641–656. https://doi.org/10.1080/02602938.2013.860950

If the methodology in the first study is valid, it shows that administrators evaluate instructors more or less highly based on statistically insignificant differences in their SETs. The second study (in the same article) found that the administrators were influenced by small fluctuations in evaluation, it was largely unintentional. The third study was better formed in terms of realistic situations, and also found that small differences in teaching evaluations have a disproportionate effect on how they are judged by administrators. In general, it calls into question the statistical validity of judgements based on SETs.

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review, 41*, 71-88. https://doi.org/10.1016/j.econedurev.2014.04.002

This article demonstrates the effect that students are unlikely to give high ratings to classes they found more demanding, even though the rigour caused them to do better in subsequent classes

Room D620, University Hall
4401 University Drive West, Lethbridge, AB T1K 3M4
www.ulfa.ca
Phone: 403 329 2578 Fax: 403 329 2113

– thus demonstrating that they did learn from the initial class they rated poorly.

Carpenter, S. K., Wilford, M. M., Kornell, N., & Mullaney, K. M. (2013). Appearances can be deceiving: Instructor fluency increases perceptions of learning without increasing actual learning. *Psychonomic Bulletin & Review, 20*(6), 1350–1356. https://doi.org/10.3758/s13423-013-0442-z

Based on a proper random experimental study at Iowa State University, students watched one of two videos of an instructor presenting a scientific concept (a fluent* instructor and a disfluent instructor performed by the same person following exactly the same script, same duration, same camera position). *Fluency encompasses the expressiveness of the instructor, use of humor, upright posture, maintaining eye contact, use of relevant gestures, not using notes. The disfluent instructor hunches over podium, reads from notes, speaks haltingly and does not maintain eye contact.

Students perceived to have learned more from the fluent instructor (results presented with high statistical significance, thus not attributable to the vagaries of sampling). The fluent instructor was also ranked significantly higher by students, who rated this instructor to be more organized, knowledgeable, prepared and effective. Though one could argue that this kind of experiment is not predictive, another experiment cited here suggests that perceptions of a 6 second silent video accurately predicts end of the semester evaluation rankings.

After watching one of the two videos students were given the text script from the presentation to restudy its content. Then they completed a test. Interestingly, actual learning did not vary significantly in relation to instructor fluency (students learned basically the same, regardless—students perceived learning to be higher with the fluent instructor, but it was not). That prompts the authors to suggest that instead of asking students whether content was delivered clearly, students should be asked whether they would be able to explain it clearly as a measure of learning (though, as they suggest, students who watched the fluent structure tend to be overconfident about what they have learned). Disfluent presentations may harm student performance not because of the delivery of content, but because students may become apathetic (stop attending, lose interest). They conclude that teaching evaluations should be corroborated with objective measures of student learning.

Clayson, D. E., & Sheffet, M. J. (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education, 28*(2), 149–160. https://doi.org/10.1177/0273475306288402

The paper addresses the question of whether student's perception of the instructor's personality influences student's evaluation of instruction. Research study compared student evaluations of instructor's personality taken at different times during the semester with end-of-semester student evaluations. The paper provides an interesting literature review, with numerous references from the past forty years supporting the relationship between student's perceptions of instructor personality and student evaluations.

The study presented in the paper was conducted in 14 sections of introductory undergraduate business courses. The study results demonstrated a statistically significant association between

Room D620, University Hall
4401 University Drive West, Lethbridge, AB T1K 3M4
www.ulfa.ca
Phone: 403 329 2578 Fax: 403 329 2113

student's perception of the instructor's personality and student's evaluation of teaching effectiveness. This relationship held across any student variable tested such as age, gender, GPA, and expectation for the class. The study showed that student's initial (within first five minutes) perception of instructor personality is significantly related to student evaluations of effectiveness and the association strengthened over time. Changes in instructor personality during the semester were shown to be significantly associated with changes in student evaluations of teaching effectiveness.
The paper argues that it is possible that student's perception of instructor personality and student's evaluation of teaching effectiveness may be largely measuring the same thing. The authors do not conclude that student evaluations should be eliminated, but that care must be taken in their use and that it must be recognized that some faculty may never receive high evaluations for reasons that may have little connection to teaching effectiveness.

DeFrain, E. (2016). An analysis of differences in non-instructional factors affecting teacher-course evaluations over time and across disciplines in the graduate college. Tucson, AZ: University of Arizona. Retrieved from http://hdl.handle.net/10150/621018

This dissertation was not read in its entirety, as it is very wide-ranging and comprehensive on the subject of SET. However, looking at the summarizing discussion indicates that such factors as class size, course level, whether it was an elective or required course, what the workload or difficulty of the class was, class standing, semester term, expected grade, discipline, and instructor gender can all play a part in how instructors are evaluated. Caution is urged when using SET.

El-Alayi, A., Hansen-Brown, A. A., & Ceynar, M. (2018). Dancing backward in high heels: Female professors experience more work demands and special favour requests, particularly from academically entitled students. *Sex Roles.* https://doi.org/10.1007/s11199-017-0872-6

Two different studies were conducted; one focused on professors' self-perception of student requests, the other on student perceptions of how they would behave with a fictional female vs male professor. Women professors did report a higher number of student requests, and felt the additional burden from them. This was noted mainly by those students who felt entitled to their education, regardless of effort. These academically entitled students also responded to the fictional scenario that they would be more likely to expect favours, and continue to ask for those favours, from a woman vs a man. Consequently, women instructors who do not cave to these request risk being penalized on student evaluations; those that do, suffer the additional time and emotional burden of the requests which may hamper their career in other ways.

Ewing, V. L., Stukas, A. A. J., & Sheehan, E. P. (2003). Student prejudice against gay male and lesbian lecturers. *Journal of Social Psychology, 143*(5), 569–579. Retrieved from http://web.csulb.edu/~djorgens/ewing.pdf

Room D620, University Hall
4401 University Drive West, Lethbridge, AB T1K 3M4
www.ulfa.ca
Phone: 403 329 2578 Fax: 403 329 2113

This article discusses was in which subtle prejudice can affect both negatively and positively – which helps bolster the argument against the validity of SETs and why they should not be used for evaluative purposes, as there are too many variables at play for a fair assessment of teaching ability. Researchers found that gay/lesbian teachers were judged more harshly when teaching well, but students rated them more leniently when their lectures were actually poorer – apparently trying to compensate and purposely not penalize someone they already felt was marginalized.

Felton, J., Koper, P. T., Michell, J., & Stinson, M. (2008). Attractiveness, easiness, and other issues: Student evaluations of professors on RateMyProfessors.com. *Assessment & Evaluation in Higher Education, 33*(1), 45-61. https://doi.org/10.1080/02602930601122803

This study expands on an initial study done by Felton in 2004. As with the previous study, correlations were found with students giving higher ratings of course quality to those classes they considered easy, and to instructors they considered attractive. Warnings about the use of SET institutionally for evaluation are given.

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), 627-630. https://doi.org/10.1037/0022-0663.86.4.627

This study found that gender role of the instructor was more important in determining SET than actual gender. Both female and male students preferred instructors who possessed both feminine and masculine characteristics, regardless of the gender of the instructor. This was particularly marked for science instructors.

Hamermesh, D. S., & Parker, A. (2005). Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review*, *24*, 369–376. https://doi.org/10.1016/j.econedurev.2004.07.013

In this article, the authors found that surprisingly, beauty matters more for men than women. The impact of beauty on evaluations was greater in lower-division courses; in these early classes, minorities, non-tenure track teachers, women, & non-native English speakers were all given lower ratings by students – and this effect persisted in the upper-level courses as well for minorities & non-native English speakers. They found female minority faculty were given lower evaluations, but not male minority faculty members. Women in general received lower evaluations than men, and minorities in general received lower evaluations than White faculty. Female minority faculty received particularly low ratings. However, male non-native English speakers received lower evaluations than female non-native English speakers; overall, though, non-native English speakers received lower evaluations than native English speakers.

UNIVERSITY OF LETHBRIDGE FACULTY ASSOCIATION

Room D620, University Hall
4401 University Drive West, Lethbridge, AB T1K 3M4
www.ulfa.ca
Phone: 403 329 2578 Fax: 403 329 2113

Hendrix, K. G. (1998). Student perceptions of the influence of race on professor credibility. *Journal of Black Studies*, *28*(6), 738–763. http://www.jstor.org/stable/2784815

This paper examines the student perceptions on (male) professor credibility (given his race, age, teaching experience, and department affiliation) at a predominantly White (student and faculty) institution. Only three professors were studied; 3 Black and 3 White – a purposeful sample. Twenty-eight students were enrolled in one of six courses taught by one of the six chosen professors. Nonparticipant observation schedule, semi-structured student interviews and professor credibility survey were used in the study. The qualitative results gathered from the three methods show that race alone would not automatically establish a professor's credibility. In addition, Black students believe strongly that Black professors work harder to earn their academic positions. However, once credibility has been established by Black professors, there is general tendency of favorable/fair attitude toward them by all students.

Kornell, N., & Hausman, H. (2016). Do the best teachers get the best ratings? *Frontiers in Psychology, 7*, 570. https://doi.org/10.3389/fpsyg.2016.00570

In this review article the authors set out to answer the question: do college students learn relatively more from teachers whom they rate highly on student evaluation forms? The study found that teachers who made their courses difficult in productive ways scored poorly on evaluation tests performed at the end of the semester, but scored well on tests conducted when the students had completed subsequent courses. Essentially, students give low marks for courses considered difficult at the time, but showed appreciation for the learning that had resulted from that experie nce. Professors who scored highly when tested at the end of the semester, and who had made the course less challenging, rated considerably lower when students were tested during subsequent semesters. Quote: "Therefore, we consider the better teachers to be the ones who contribute the most to learning in subsequent courses… We refer to this kind of generalizable knowledge as "deep learning."" The authors conclude that instructors who created the most meaningful learning were ones who did not necessarily rate well at the end of the first course, but rated higher on successive courses.

Lazos, S. R. (2012). Are student teaching evaluations holding back women and minorities? The perils of "doing" gender and race in the classroom. In G. G. y Muhs, Y. F. Niemann, C. G. Gonzalez, & A. P. Harris (Eds.), *Presumed Incompetent: The intersections of race and class for women in academia* (pp. 164–185). Boulder, CO: University Press of Colorado. http://darius.uleth.ca/record=b2178457~S1

This chapter summarizes much of the recent research on the validity of SETs, especially as they pertain to women and minorities. It echoes many of the themes in other articles such as the difference in gender expectations and the repercussions for those who don't fit those norms. It cites the grade inflation (and grade expectations) that predict positive evaluations; the fact that deeper learning may engender lower evaluations; and that ultimately, ratings are more subjective than objective and are swayed by non-teaching qualities as beauty and charisma. A

good summary of why at the very least, the validity and reliability of SETs should be questioned. There is also reference to an article in which women and minorities gain tenure at lower rates and earn less in merit increases.

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. https://doi.org/10.1016/j.stueduc.2016.12.004

Linse is critical of the scholarship critical of SETs, but does acknowledge the possibility of bias, especially "due to race, ethnicity, or culture" (p. 98). Her assessment of the lack of bias against women is based on articles that are over 20 years old. Despite her support of SETs, she has a useful set of recommendations for administrators and faculty members serving on STP committees for their use.

Macfadyen, L. P., Dawson, S., Prest, S., & Gašević, D. (2015). Whose feedback? A multilevel analysis of student completion of end-of-term teaching evaluations. *Assessment & Evaluation in Higher Education, 41*(6), 821–839. https://doi.org/10.1080/02602938.2015.1044421

While this article does not specifically address the issue of instructor gender or minority bias in SET, it does discuss the issue of response biases in SET more generally and its effects based on student gender, age, grade, term, and whether the class is part of their major. While the authors do not advocate doing away with SET, they do provide suggestions in increasing student response rate in order to get a more representative sample as well as educating faculty on the limitations of the instrument.

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.https://doi.org/10.1007/s10755-014-9313-4

In an online course, one male and one female instructor were each assigned two different discussion groups, one which was aware of the instructor's actual gender, and the other which believed the instructor was the opposite gender. So of the four groups, one was taught by a male and the students believed it was a male; one was taught by a male, but students believed it was a female; one was taught by a woman, and students believed it was a woman; and one was taught by a woman, but students believed it was a male. Students rated the instructors they believed to be male significantly higher than those who believed they had a female instructor, regardless of the instructor's actual gender. The same instructor received different ratings based on their perceived gender, presumably because of the incongruence between perceived congruence of gender traits and what the students believed was the gender of the instructor. The authors found that men are afforded an automatic credibility, and that female instructors must work harder than male instructors to receive comparable ratings. They urge institutions to reassess the use of SET, as it may systematically disadvantage women in academia.

UNIVERSITY OF LETHBRIDGE FACULTY ASSOCIATION

Room D620, University Hall
4401 University Drive West, Lethbridge, AB T1K 3M4
www.ulfa.ca
Phone: 403 329 2578 Fax: 403 329 2113

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187-1197. https://doi.org/10.1037/0003-066X.52.11.1187

The authors make the case in support of student evaluations of teaching effectiveness. They point to studies involving multiple sections of the same course taught by different instructors that show statistically significant correlations between students' achievements and components of the student evaluation of teaching effectiveness. They also point to studies showing general agreement between student evaluations and the evaluations of trained external observers. Nevertheless, the authors reiterate that no single measure of teaching effectiveness is adequate taken on its own, but point to the difficulty in finding general support for other indicators. The authors discuss several factors that are associated with student evaluations, but these are limited to such factors as class size, prior subject interest, expected grades, workload, reason for taking the course, instructor expressiveness. The potential effect on student evaluations of such factors as instructor age, gender, race, etc. are not discussed in the paper.

Martin, L. L. (2016). Gender, teaching evaluations, and professional success in political science. *PS: Political Science & Politics, 49*(2), 313-319. https://doi.org/10.1017/S1049096516000275

Examines data from large political science classes and finds female instructors receive significantly lower evaluations than male instructors. Summarizes much previous research including gender stereotyping, gaming of SETs, and the general subjectivity of SETs.

Mayer, R. E., Sobko, K., & Mautone, P. D. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology*, *95*(2), 419–425. https://doi.org/10.1037/0022-0663.95.2.419

This study tested students' evaluations of computer-based instruction; some of them listened to narration with non-accented speech, and some with a foreign accent (and some with a computer-generated voice, which is not part of our analysis). Students rated the instructors with the foreign accent lower than that with no accent. While the point of this research was to test the best voice to have students learn online, it still shows that something as tangentially related to teaching ability as the instructor's voice can affect student evaluations.

Mengel, F., Sauermann, J., & Zolitz, U. (2017). *Gender bias in teaching evaluations* (IZA Discussion Paper No. 11000). Retrieved from: https://ideas.repec.org/p/iza/izadps/dp11000.html

A large-scale study done with detailed student data from a school of Business & Economics in the Netherlands, this study showed that students rated women professors lower than men. Male students rated women even lower than female students rated women, though both rated women lower than men, and junior women faculty as well as those teaching in math-related courses. Even the teaching materials were rated lower, even though all sections received the same courseware and readings. Course performance and self-reported study time was not

Room D620, University Hall
4401 University Drive West, Lethbridge, AB T1K 3M4
www.ulfa.ca
Phone: 403 329 2578 Fax: 403 329 2113

affected by the gender of the instructor. The authors muse about implications beyond STP process, including dedicating more time to improving teaching (taking time away from research), and losing confidence in one's own teaching to the point of ceasing one's career in academia.

Mitchell, K. M. W. & Martin, J. (2018). Gender bias in student evaluations. *PS: Political Science & Politics,* 1-5. https://doi.org/10.1017/S104909651800001X

This article contends that the use of SETs in hiring and promotion activities constitutes discrimination due to the inherent biases. Language used in both official school-administered evaluations as well as *Rate My Professor* indicated gendered differences in the language used to evaluate instructors, as well as the criteria used to evaluate them. The authors also taught identical sections of the same online course, but the woman was consistently rated lower than the man, despite slightly higher grades in the sections taught by a woman.

Merritt, D. (2008). Bias, the brain, and student evaluations of teaching. *St John's Law Review*, *82*(1), 235–287. https://doi.org/10.2139/ssrn.963196

This review article discusses general biases in SET. The references are more general and some solutions of how to deal with student evaluations in STP situations are presented. It proposes new evaluation methods to avoid biases and be more accurately focused on actual teaching abilities. It refers to studies in which judgements on teaching ability are made within a few minutes of meeting an instructor, and don't change over the course of the class. It also talks about the problem of grade inflation as teachers try to 'game' the evaluations (students rate easier classes, better). The author argues that it's been proven that SET are biased, but not in a way that can be simplistically stated but rather based on nonverbal behaviours (many of which, it can be argued, are actually rooted in gender or race) including body language, voice, and appearance. Things like acting classes can improve an instructor's rating, but is that really making them a more effective teacher? Communication is important, but it figures disproportionately in students' evaluations. There are also different standards for women than men of what students expect from them (based on gender norms) and how they evaluate instructors based on their perceived fit with those norms.

Nasser-Abu Alhija, F. (2017). Teaching in higher education: Good teaching through students' lens. *Studies in Educational Evaluation, 54*, 4-12. https://doi.org/10.1016/j.stueduc.2016.10.006

The author posits that students and instructors may have different ideas of what constitutes 'good teaching', and that that definition varies among age/year of student along with gender, field of study, and institution type. Students were asked their evaluation of the importance of five dimensions of teaching: goals to be achieved; long-term student development; teaching methods and characteristics; relationships with students, and assessment. Unsurprisingly, younger undergraduates were most concerned with assessments and grades, while older students were more aware of the importance of long-term student development. This article

Room D620, University Hall
4401 University Drive West, Lethbridge, AB T1K 3M4
www.ulfa.ca
Phone: 403 329 2578 Fax: 403 329 2113

goes to the argument that what we are measuring with evaluations is more about student satisfaction with the particular course/instructor, rather than actual learning taking place. The author concludes, "…student ratings of teaching quality should be considered with caution, for formative (teaching improvement) and summative purposes, especially for high stakes use." (p. 11).

Pittman, C. 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. https://doi.org/10.1177/0092055X10370120

This qualitative study explores the classroom experiences of women of colour (Black, Asian & Latina). The research found that women of colour were challenged almost exclusively by White males in their teaching activities. They face many of the same issues that women faculty face but it's multiplied by race. They also face increased isolation as there are few in the same position (there are other women and other minorities, but not that many of both). Minority scholarship is not seen as valid and their authority in any topic is not recognized. There were four themes that emerged: their authority was challenged; their competency questioned; their scholarly expertise was disrespected; and they felt threatened and intimidated (both physically and in job security). It refers to three other studies showing female minorities receive lower student evaluations. Instructors feel this to be true, leading to worries about merit increases and STP processes.

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), 137-152. https://doi.org/10.1037/a0019865

Studying a large sample of data from RateMyProfessors.com, White faculty were rated better than non-White faculty. Black faculty, and especially Black men, fared especially poorly; and although Black faculty were rated as easier, this did not correspond to a high rating in overall quality as it did with White and Latino faculty. Asian faculty were also rated lower overall than White and Latino. Racialized instructors were also rated lower in terms of helpfulness and clarity.

Rubin, D. L., & Smith, K. A. (1990). Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants. International Journal of Intercultural Relations, 14, 337–353. https://doi.org/10.1016/0147-1767(90)90019-S

This study found that the stronger the students believed an instructor's accent was (whether or not it was actually very strong), the students rated them as poorer instructors. The level of accented speech did not affect student comprehension of the material, just their perception of their level of comprehension.

Rubin, D., Angletti, S., Caudill, S., Harzog, B., Kilgore, M., Li, L., … Yeardley, J. (1995). Effects of language and race on undergraduates' perceptions of language and attitude in higher

Room D620, University Hall
4401 University Drive West, Lethbridge, AB T1K 3M4
www.ulfa.ca
Phone: 403 329 2578 Fax: 403 329 2113

education. *Paper presented at the Annual Meeting of the International Communication Association* (pp. 2–8). Alberquerque, NM. Retrieved from http://files.eric.ed.gov/fulltext/ED384932.pdf

In another variation on the previous study, the researchers found that accent did not affect comprehension; rather, a clearly organized, not too dense lecture was more effective. Physical attractiveness affected students' perception of teaching effectiveness. Those who spoke with less of an accent were rated more highly than those with a thicker accent.

Sanchez, C. A., & Khan, S. (2016). Instructor accents in online education and their effect on learning and attitudes. *Journal of Computer Assisted Learning, 32*, 494–502. https://doi.org/10.1111/jcal.12149

The study discussed in the paper is quite limited in scope. A group of undergraduate volunteers were presented with a six-minute video lesson in object-oriented computer programming, narrated by either a native English speaker or a non-native English speaker. Student learning and student evaluation of the instructor and the quality of the lesson were measured immediately afterward.

Differences in narration did not have any discernible effect on student learning or on student attitudes towards the subject matter or instructional method. There was a significant difference in attitudes towards the instructor between the two groups. Students rated the native English speaker significantly higher when asked to rate the quality of the instructor. There was also an interesting, perhaps concerning, difference discovered between the two groups regarding factors affecting change in attitudes towards subject matter or instructional method. In the native English speaking group, changes in attitudes were predicted by learning, whereas in the non-native English speaking group, changes in attitudes were predicted by attitudes towards the instructor.

Despite the limitations of the study, it does highlight the potential challenges faced by non-native English speaking instructors, by examining the disconnect between student learning and student evaluation of teaching effectiveness.

Smith, B. P., & Hawkins, B. (2011). Examining student evaluations of Black college faculty: Does race matter? *The Journal of Negro Education, 80*(2), 149-162. http://www.jstor.org/stable/41341117

This study, a version of which was also published in 2007, found that Black faculty received lower ratings than any other minority or White faculty on overall value of the course and teaching ability, despite being rated similarly on individual factors as White and other minority faculty.

Smith, B. P. & Johnson-Bailey, J. (2011). Student ratings of teaching effectiveness: Implications for non-White women in the academy. *The Negro Educational Review, 62/63*(1-4), 115-140. https://search-proquest-com.ezproxy.uleth.ca/docview/940916203?accountid=12063

Room D620, University Hall
4401 University Drive West, Lethbridge, AB T1K 3M4
www.ulfa.ca
Phone: 403 329 2578 Fax: 403 329 2113

Limiting the data from Smith's previous study to just female faculty members, the results here also show that Black women faculty are rated lowest overall, followed by other minorities, and White women faculty being rated the highest. Again, this despite data showing their actual teaching effectiveness being similar.

Smith, G., & Anderson, K. J. (2005). Students' ratings of professors: The teaching style contingency for Latino/a professors. *Journal of Latinos and Education, 4*(2), 115-136. https://doi.org/10.1207/s1532771xjle0402_4

This study, unlike Anderson's other study on Latina professors, did find an effect between teaching style and student ratings, where warmth was expected and rewarded from Latino/a faculty and penalized when absent.

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. https://doi.org/10.3102/0034654313496870

An extensive overview of literature on SETs since 2000. It concludes with the thought that SETs should be used very cautiously for evaluative purposes because the studies are so inconclusive in their differing results. The instruments used vary greatly by institution and even within institution, and thus are extremely difficult to compare results from; furthermore, evaluators differ on what constitutes good teaching. Questionnaire design can influence outcome a great deal. One point raised was that institutions are often using these and evaluated on these as an institution – as such, it's really in the institution's best interest to ensure their validity. On the question of bias, the article notes that the overview of studies points out potential bias in the area of course discipline and instructor characteristics such as sexual orientation, gender, and race – though overall, the authors felt the bias to be small. Still, they raise questions on the overall validity of SETs and as such advise caution on their use for any type of evaluative purpose.

Stark, P. B., & Freishtat, R. (2014). An Evaluation of Course Evaluations. *ScienceOpen*, 1–26. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

The authors of this review article suggest that teaching evaluations do not measure teaching effectiveness. Student responses shed light on certain aspects of teaching, but not on teaching effectiveness as such. Still, it is unclear what they ultimately reflect (e.g. "clarity" may be confounded with "difficulty"—an instructor teaching with clarity, but teaching difficult material is at a disadvantage). Further, the response rate greatly affects the scores that professors get in evaluations; students are more likely to fill out evaluations when they dislike a course/instructor. Analysis of data from U of California, Berkeley suggests that scores are highly correlated with students' grade expectations and "enjoyment" scores. More disturbing, scores can be predicted from students' reaction to 30 seconds of a silent video (physical attractiveness is a factor). Gender, ethnicity, and age also are factors. Questions about course design and effectiveness are

Room D620, University Hall
4401 University Drive West, Lethbridge, AB T1K 3M4
www.ulfa.ca
Phone: 403 329 2578 Fax: 403 329 2113

most influenced by factors unrelated to learning. In written comments, students tend to focus on difficulty of the material or workload, not on pedagogy or teaching effectiveness.
The authors suggest other approaches to evaluating teaching effectiveness, such as looking at samples of student work, submission of teaching portfolios by faculty, and peer evaluation by other colleagues (which would need to become standard practice in order to control for bias and avoid the implication that something is 'wrong', resulting in the peer review).

Stonebraker, R. J., & Stone, G. S. (2015). Too old to teach? The effect of age on college and university professors. *Research in Higher Education, 56*(8), 793–812. https://doi.org/10.1007/s11162-015-9374-y

Another study which utilizes RateMyProfessors.com to analyze student evaluations of instructors, this time based on age. Results show that age does affect students' perception of teaching effectiveness, starting in the mid-forties and remaining constant. While the authors contend that the age effect can be offset by maintaining one's appearance or 'hotness' level and marking easily, clearly these are not acceptable ways of mitigating the unfair bias against an aging instructor.

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. http://dx.doi.org/10.1016/j.stueduc.2016.08.007

An extremely comprehensive meta-analysis which debunks previously highly-cited publications stating a correlation between SET & student learning. The authors conclude that the correlation between the two is nil, and that these ratings should be given little to no weighting for evaluative purposes as they are more likely to measure student *satisfaction* than student *learning.*

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 23*(2), 191-211. https://doi.org/10.1080/0260293980230207

This review paper goes through the history of SET and summarizes much of the literature to 1998. It notes that while SET feedback can help improve instruction, other feedback is also required. Also noted are many of the methodological and validity concerns with the instruments. It goes through the literature citing the ways in which SET can be effected that have nothing to do with the instructor, such as student background, timing of evaluations, class time, class level, class size, subject area, and workload. It cites the tendency for professors to be rated more highly than teaching assistants and the different personality characteristics which correlate with higher and lower ratings.

Wagner, N., Rieger, M., & Voorvelt, K. (2016). Gender, ethnicity and teaching evaluations: Evidence from mixed teaching teams. *Economics of Education Review, 54*, 79–94. https://doi.org/10.1016/j.econedurev.2016.06.004

UNIVERSITY OF LETHBRIDGE FACULTY ASSOCIATION

Room D620, University Hall
4401 University Drive West, Lethbridge, AB T1K 3M4
www.ulfa.ca
Phone: 403 329 2578 Fax: 403 329 2113

This study analyzes a dataset of evaluations from a five year period. The researchers tried to control for all sorts of variables which could come into play (tenure, new hires, age, course leadership, outliers) and still found women's evaluations to be statistically significantly lower than men's. While they did find a positive relationship between research quality and evaluations, it was not enough to mitigate the gender effect ("a female teacher would need almost five A publications in the year prior to the evaluation to offset the direct negative gender effect on teaching evaluations" (p. 88).) They also discovered that mixed gender teaching teams were not negatively impacted. They did not find a similar negative correlation with evaluations and ethnicity.

Weinberg, B. A., Fleisher, B. M., & Hasimoto, M. (2007). *Evaluating methods for evaluating instruction: The case of higher education* (NBER Working Paper No. 12844). Retrieved from National Bureau of Economic Research: http://www.nber.org/papers/w12844

This study showed that students give good evaluations to those who are easier markers; evaluations are influenced by current grades, but not by an actual measure of the learning taken place. Thus, ratings can be manipulated by giving better grades. Some evidence was also found in this study that women and foreign-born instructors were also given lower SET. Other, more optimal ways of evaluating teaching are provided.

Yunker, P. J., & Yunker, J. A. (2003). Are student evaluations of teaching valid? Evidence from an analytical business core course. *Journal of Education for Business*, *78*(6), 313–317. https://doi.org/10.1080/08832320309598619

This article notes that students taking lower level, intro courses – especially if those courses are required – tend to rate those instructors more poorly. This is more a factor of the student's interest than teaching effectiveness, but often these courses are disproportionately taught by women. In this specific study, the authors found that students from a highly-rated intro course (prerequisite) actually did worse in the subsequent class. This presumes that students learned less in the introductory course, but rated the instructor more highly (possibly because it was easier). They also found that students with higher grades tended to give higher evaluations (thus leading to easier classes getting higher evaluations).

*Thank you to GEDC committee members for 2016-2017 Andrea Cuellar, James Graham, Bente Hansen, John Sheriff, D. Andrew Stewart, Kien Tran, & Kelly Williams-Whitt for initial discussions and article evaluations, and to the members from 2017-2018 Robert Benkoczi, Andrea Cuellar, Bente Hansen, Carolyn Hodes, Sheila McManus, & John Sheriff for further discussion on the final statement.

Nicole Eva
Chair, Gender, Equity and Diversity Committee
University of Lethbridge Faculty Association