



BAYLOR
UNIVERSITY

ACADEMY FOR TEACHING AND LEARNING

Justification, Critique, and Use of Course Evaluations: A Brief Annotated List of Resources

Prepared by Christopher Richmann, Assistant Director¹

Scholarly Articles:

Benton, S.L. & Cashin, W.E. (2012). Student Ratings of Teaching: A Summary of Research and Literature. *IDEA Paper 50*. <http://ideaedu.org/wp-content/uploads/2014/11/idea-paper-50.pdf>

A meta-analysis that supports the use of student evaluations of teaching. The report concludes: “In general, student ratings tend to be statistically reliable, valid, and relatively free from bias or the need for control, perhaps more so than any other data used for faculty evaluation” (12). Some suspicion may be in order, since this is a study done on behalf of an organization that produces an evaluation (IDEA). The study fails to elaborate on what “important aspects of teaching...students are not competent to rate” (2) and may subtly downplay evidence that challenges its conclusions (referring to correlations as “low” [8] or “quite low” [10] when data did not fit the argument, and “correlated positively” when comparable statistics support their argument [3]). Still, the study offers useful suggestions for increasing ratings (4) and cautions that SET should be used carefully and judiciously.

Boring, A., Ottoboni, K., and Stark, P. (2016). Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness. *SOR-EDU*. <https://www.scienceopen.com/document/read?vid=818d8ec0-5908-47d8-86b4-5dc38f04b23e>

Through a combination of data from a natural experiment and a controlled experiment, the authors find that student evaluations of teaching are biased against female instructors, such that, for instance, ineffective male instructors receive higher ratings than effective female instructors. This bias even extends to student ratings of objective aspects of teaching, such as how promptly assignments were returned. The authors also find that student evaluations are more sensitive to gender bias and grade expectations than they are to teaching effectiveness, and that gender bias, in particular, depends on so many factors (discipline, student gender, etc.) that it is impossible to adjust for the bias. The authors contend that other studies finding no association between gender and student evaluations fail to control for other variables, such as teaching effectiveness.

¹ Updated July 30, 2017

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.

<http://onlinelibrary.wiley.com/doi/10.1002/tl.10001/epdf>

This article teaches faculty how to advocate for appropriate and constructive use of evaluations by creating a “narrative” to include with your teaching or promotion portfolio. Constructing a narrative can help in two ways: protecting against misinterpretation, misuse and bad data from student evaluations, and helping you see how improving your teaching can improve your ratings. Franklin writes, “Submitting a well-written, well reasoned narrative discussing your students’ evaluations of your teaching (that is, ratings) is an opportunity to improve the odds that your reviewers will consider students’ opinions in the full context of the complex factors that shaped them” (85). Faculty reviewers should be led to understand that they (not the students via evaluations) are the true evaluators and that student evaluations are “data for review by colleagues, not canned, ready-made evaluations on which to base personnel decisions” (85). But because faculty are prone to misinterpret or misuse this data, the instructor must put these numbers in proper perspective. Franklin discusses important mitigating factors (such as class size) and also gives some technical advice, for instance, on how to calculating margin of error and “confidence interval,” which can be especially helpful if your ratings are slightly lower when compared to others in your department.

Smith, G. and Anderson K.G. (2005). Students’ Ratings of Professors: The Teaching Style Contingency for Latino/a Professors. *Journal of Latinos and Education* 4, 115-136. http://www.tandfonline.com/doi/abs/10.1207/s1532771xjle0402_4

A study, based on student interaction with a hypothetical course syllabus, dealt with three variables: teaching style (lenient/strict), gender, and ethnicity (Latino/a/Anglo). The study did not find significant differences in the way students viewed men and women. However, the study concluded that Latino professors (men and women) received low marks when they were perceived as strict and high marks when they were perceived as lenient. These associations were not present for white men or white women. The authors conclude that this indicates a “high cost for Latino/a professors who led with a strict teaching style.”

Stark, P.B., & Freishtat, R. (2014). *An Evaluation of Course Evaluations*. <https://www.stat.berkeley.edu/~stark/Preprints/evaluations14.pdf>

Summary of research. Frank and clear: student evaluations of teaching do not measure teaching effectiveness. Quantitative data from teaching evaluations should not be averaged, since that is misleading. Evaluations do give useful insight into students’ experiences and some practical issues related to teaching effectiveness, like audibility, legibility, and availability (cf. pp. 7-8). If used in professional considerations, student evaluations should be balanced with other information, especially faculty peer observation.

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://www.sciencedirect.com/science/article/pii/S0191491X16300323>

A critique of previous meta-analyses, with a history of scholarship on multi-section studies. The study finds that previous meta-analyses had many methodological flaws, and that their claimed correlations between ratings and learning are the result of small sample bias and publication bias. The authors' own meta-analysis shows low correlation—especially when controlled for small study size bias—and no correlation when adjustments of randomization and accounting for prior knowledge/ability are made. They conclude that “multisection studies do not support the claim that students learn more from highly rated professors” (39).

Williams, W.M, and Ceci, S.J. (1997). “How'm I doing?” Problems with Students Ratings of Instructors and Courses. *Change: The Magazine of Higher Learning* 29 (5), 12-23. <http://www.tandfonline.com/doi/abs/10.1080/00091389709602331>

In a “natural” experiment that controlled for most possible variables, an instructor increased the level of enthusiasm (voice pitch, tenor, and gestures) in a class compared to the same class the previous semester. The authors found that level of enthusiasm was significantly linked to improved student ratings on not only enthusiasm of the instructor, but student ratings of the instructor's tolerance of viewpoints, accessibility, and organization—all which objectively remained constant. Level of enthusiasm was also linked to student ratings of perceptions of amount learned, the quality of the text used, the fairness of grading, and the clarity of course goals—all of which, again, objectively remained constant. The authors comment on this “halo-effect,” but especially point out “the magnitude of changes in students' evaluations due to a content-free stylistic change by the instructor” (22). They recommend that if institutions persist in using student evaluations for promotion decisions, faculty should be trained in techniques to enhance their student ratings.

Website Resources:

Vanderbilt University's Tips - <https://cft.vanderbilt.edu/guides-sub-pages/student-evaluations/>

An easy-to-navigate set of points regarding student evaluations. “Talking with Students about Evaluations” recommends communicating to students the value you as instructor place on evaluations and encouraging honest and specific feedback. “Making Sense of Evaluation Feedback” guides instructors to look for patterns and trends, suggests that most negative evaluations revolve around in-class communication and course organization, and advises taking context into account—for instance small classes, electives, and courses in which students do well tend to get higher ratings. Extra resources (external links) are provided on “Interpreting Student Evaluations” and “Summaries of Research on Student Evaluations,” which provides many resources on potential gender bias in student evaluations.

University of Michigan – Tip sheet for students before evaluation

http://crlt.umich.edu/sites/default/files/resource_files/Course%20Evaluation%20Guidance%20One-Pager.pdf

A simple and useful one-page tool for instructors to give to students (or discuss with students) that directs students in how to provide the most useful feedback/written comments by being more specific, concrete, and constructive. Using such a tool will not only likely help instructors get more useful feedback, but it will also communicate to students that the instructor values student feedback.

University of Indiana – “New Measures for Course Evaluation”

<https://www.indiana.edu/~tedfrick/TALQ.pdf>

The authors contend that traditional student evaluations offer limited guidance for improving courses. The proposed alternative is called a “Teaching and Learning Quality” scale. Most importantly, this instrument measures the students’ perception of the instructor’s use of “First Principles of Instruction” (FP) and the students’ own Academic Learning Time (ALT). FPs focus in particular on authentic tasks, connections to what students already know, demonstrations, opportunities for students to test new knowledge with instructor feedback, and integration of knowledge into students’ personal lives. ALT “refers to repeated successful student engagement in learning activities that are relevant to student goals” (2). Student mastery is positively correlated to both FP and ALT, although mastery has a higher correlation to ALT than FP. The biggest payoff, however, is that students are 5.2 times more likely to achieve high mastery and 26 times less likely to achieve low mastery when students indicate that FP and ALT had occurred.

Dartmouth University – End of Term Evaluations

<http://dcal.dartmouth.edu/resources/course-assessments-and-evaluations/end-term-evaluations>

Some of the information on this website is specific to Dartmouth’s student evaluation process. More general information here includes helpful reminders on student biases and the difference between median and mean. The site also provides elaborations of typical evaluation questions, which gives insight into what students understand by various prompts and suggests ways instructors can improve these ratings. For example, a typical “rate overall teaching effectiveness” prompt is often a measure of whether students thought the instructor was invested in the student’s success.

Stanford University – Using Student Evaluations to Improve Teaching

https://web.stanford.edu/dept/CTL/Newsletter/student_evaluations.pdf

This short essay provides specific tips about organizing student comments into useful categories so that patterns can emerge. For instance, what are the most common comments for improving a class among students who gave high global ratings, or those who gave low global ratings? The bottom line is that apparent contradictions in written responses should not lead instructors to dismiss such comments as meaningless.