Bibliography (Brief) on Student Evaluations and Bias

--compiled by Julie Sievers, Center for Teaching, Learning, and Scholarship, Jan/Feb 2020

How to control your reading time

I am recommending that you read or skim the following. I have included some additional options for reading, but suggest that you start with these. *You can read, or at least skim, all of the sources I outline below in about 1-1.5 hrs, I think.*

1. In the "Summarizing the State of the Literature" section below: items 1 and 2.
2. In the "On Bias, Specifically" section below: item 1
3. In the "Guidance for Evaluators" section, items 1 and 2.

Summarizing the state of the literature

We can't talk about the research on bias and student evaluations without understanding the basic research on student evaluations themselves. There are *thousands* of research articles about student evaluations of teaching, however. It is hard to talk about "the research on student evals," as if it is a consistent body of literature, because it's so huge and messy. But I am going to point you to a few synthesis accounts that are: a) widely known and respected; and b) taken together, give you a fairly quick sense of the contours of this literature.

A few "heads up" items before we begin:

- First, when we talk about "the research on student ratings," we should clarify what we are talking about. Researchers on this topic have generally attempted to identify a relationship between two things: the quantitative scores on the global evaluation question ("Overall, this instructor was excellent" / "Overall, this course was excellent"), and student learning, as measured by final exam performance or final course grade or performance in a subsequent course. That is, early on in this research, student ratings were studied to determine whether they could serve as a proxy measure for teaching effectiveness, which was equated with student learning, as measured in these ways. These are now seen as a problematic set of assumptions. And although the qualitative feedback and the criterion-specific questions (for example: "The class was well organized" or "Feedback was provided in a constructive fashion") might prove useful for identifying whether an instructor teaches in ways that match a university's or department's ideas of good teaching (less problematic), the published research literature rarely looks at these components of ratings. Notably, almost none of the research addresses the things that we say we focus on the most at SU in our evaluation processes: the written comments.

- Second, the research literature has spanned over 80 years. People trying to summarize this literature often cite this as a good thing -- as an argument for taking this research seriously, and for seeing it as having results that we should listen to. (For example, Linse, below, cites positively that this is a "vast literature" and "the most researched topic in higher education" (95)). But . . . higher education has changed a *lot* over 80 years. It wasn't until I read Linda Nilson's piece -- included below -- that I realized that making

arguments about the validity of student ratings based on a meta-analysis published in 1981 (one of the seminal pieces on which many of the ratings-are-valid arguments are based - and included below), is questionable. A *lot* has changed about our students, professors, institutions, social contexts, courses, etc., since the 1960s and 70s, when the studies summarized in the 1981 meta-analysis were done. So - it's important to understand this: the research doesn't just vary in what researchers are saying at any given moment in time; some of the differences in the research are a product of *when* the studies were done. This literature is not monolithic, and a lot of it is concerningly dated.

1) **University of Michigan's Center for Research on Learning and Teaching,** "Frequently Asked Questions about Student Ratings: Summary of Research Findings"
    - A fairly concise summary of major contours of the research.

2) **Betsy Barre's 2015 and 2018 literature reviews** - both published while she was at the Rice University Center for Teaching Excellence. (Barre is now Executive Director of the Wake Forest Teaching and Learning Collaborative).
    - 2015, "Do Student Evaluations of Teaching Really Get an 'F'"?
        - This piece is widely known and shared, and it is helpful for trying to navigate between what we hear in the *news* about student ratings, which tends to emphasize the most outrageous problems with them, and what the research literature actually says, which is generally more nuanced and cautious.
        - She then provides some helpful summaries of key takeaways from much of the literature.
        - She ultimately argues that they are an imperfect but still useful measure of teaching effectiveness.
    - 2018, "Research on Student Ratings Continues to Evolve. We Should, Too."
        - This piece *reverses* some of her earlier conclusions, based on new research.
        - I suggest reading these back to back, for a sense of how this discussion jas fairly quickly changed.
        - Her final takeaways are helpful and worth a look.

    - *Alternate 2018 option* - Prefer to get your lit digest in audio format? Try Doug McKee and Edward O'Neill's *Teach Better* podcast interview with Betsy Barre, which overviews the student evaluation literature: http://teachbetter.co/blog/2018/03/22/tbp-episode-72/

3) Optional: Spooren, Pieter, Bert Brockx, and Dimitri Mortelmans. (2013). "On the validity of student evaluation of teaching: The state of the art," in *Review of Educational Research* 83 (4): 598-642.
    - Conclusion: "This review of the state of the art in the literature has shown that the utility and validity ascribed to SET should continue to be called into question. Next to some, although much-researched, topics such as the dimensionality debate and the bias question, new research lines are delineated (i.e., the utility of online SET, teacher personal characteristics affecting SET). Our systematic use of the meta-validity framework of Onwuegbuzie et al. (2009), however, shows that many types of validity of SET remain at stake. Because conclusive evidence has not been found yet, such evaluations should be considered fragile, as important stakeholders (i.e., the subjects of evaluations and their educational performance) are often judged according to indicators of effective teaching (in some cases, a single indicator), the value of which continues to be contested in the research literature."

4) *Optional:* Uttl, Bob, Carmela A. White, and Daniela Wong Gonzalez. (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.
    - Key claims: a) Students do not learn more from professors with higher student evaluation of teaching (SET) ratings; b) Previous meta-analyses of SET/learning correlations in multisection studies are not interpretable; c) Re-analyses of previous meta-analyses of multisection studies indicate that SET ratings explain at most 1% of variability in measures of student learning; d) New meta-analyses of multisection studies show that SET ratings are unrelated to student learning.

5) *Optional:* Nilson, Linda. (2012). "Time to Raise Questions about Student Ratings," in *To Improve the Academy: The Journal of Educational Development* 31: 213-227.
    - Nilson begins this piece by pointing out that "the bulk of research on validity of student ratings and their contaminating biases was conducted in the 1970s and 1980s, when a very different generation of students filled classrooms." She proceeds to look past the findings into the methodologies and their assumptions.
    - Nilson concludes by going back to the start of all of this. "The academy has to decide what 'teaching effectiveness' really means. If it means student learning, as Cashin (1988) defined it, then institutions should assess it using measures that reflect learning, not student ratings."

6) *Optional - one of the original, widely influential, pieces arguing for a positive correlation between student ratings and student learning:* Cohen, Peter A. (1981). "Student Ratings of Instruction and Student Achievement: A Meta-Analysis of Multisection Validity Studies." *Review of Educational Research* 51(3): 281–309.

On Bias, Specifically

The overviews above will talk about bias. But, perhaps you will be surprised to realize, bias isn't actually the biggest concern in most of the literature. Overall, the research studies, taken together, seem to make a few key suggestions about bias:

- Bias does shape student ratings, but a significant number of the most frequently-cited studies suggest that the things that have been shown to bias them the *most* are not aspects of instructor identity but rather other factors, such as whether the course is required or elective, class size, and discipline. So, it's important to keep these other issues on our list of contextual factors to watch when interpreting student ratings and comments.
- As a result, some major voices have suggested that there are minimal *quantitative* effects of instructor identity on the *global* question data. Keep in mind, however, that these voices are often drawing on old research, including the 1981 meta-study, to make claims about what "the research" says. Also a factor: over the 80 years of this literature, the experiences of faculty of color and other underrepresented groups were infrequently studied, precisely because of their minoritized status in academia.
- Instructor identity does influence student ratings, however. According to Doug Holton's synthesis of this literature (cited below), "while research recognizes course evaluations are an imperfect measure, the literature indicates many of the problems with course evaluations are

unevenly distributed across men and women and across white faculty and faculty of color." (See his extensive bibliography of research on gender and racial bias and student evaluations, below.)

- Additionally, there is a significant body of research about bias, generally -- not specific to student evaluations. You have already read about some of this research in *An Inclusive Academy: Achieving Diversity and Excellence,* which synthesizes quite a lot of it. This research has shown that implicit biases affect evaluators in all sorts of evaluation situations. That necessarily includes our students when they play an evaluative role.
- While most researchers have tried to isolate a specific characteristic--instructor gender, or instructor sexuality --and then have studied how it affects student evaluations, it's important to keep in mind that many faculty live and work at the intersections of multiple identities about which students have implicit biases: they are faculty of color *and* women *and* queer, for example, and they may also be teaching required courses or large classes -- all factors that can elicit biased ratings.
- At the same time, as Betsy Barre said recently, worrying about the effect of bias is perhaps less important than realizing that "the house is on fire." The *biggest* problems we should be worrying about, she argues, are that: a) student ratings are such a messy, noisy measure of teaching effectiveness that they should never be used alone; and b) student ratings should never be used to compare faculty to one another.

1. Mengel, Friederike, Jan Sauermann, and Ulf Zolitz. (2018). "Gender Bias in Teaching Evaluations." *Journal of the European Economic Association* 17(2): 535-566.
    - I am including this as the recommended bias-specific reading because it's the one that Barre references, in her 2018 literature review piece, when she suggests that in recent years "although questions about validity never left the scene, concerns about potential biases against marginalized faculty, and particularly women, took center stage. . . . Yet, surprisingly, most of the research on gender has produced mixed results, with many studies suggesting no significant bias. Recent studies have tried to challenge this result, but as with most recent validity studies, they were often poorly designed and rarely addressed the literature [on student ratings]. Yet the soon to be published work of Friederike Mengel, Jan Sauermann, and Ulf Zolitz has been a notable exception."
    - This is a research article, not a summary or meta-analysis, so if you don't want to get into the methodological weeds, skim sections 1, 2, and 4.)

2. *Optional:* I am linking a fairly complete and up-to-date bibliography of research around gender and racial bias and student ratings - compiled by Mirya Hollman (Tulane), Ellen Key (Appalachian State) and Rebecca Kreitzer (UNC-Chapel Hill). It contains links to a Google folder that contains PDFs of each article listed in the biblio.
    - (*Correction:* This document was mis-attributed in the original version of this bibliography.)
    - Update 3/9/21: Kreitzer has since published some of her work on this topic. Kreitzer, Rebecca J. and Jennie Sweet-Cushman. (2021). "Evaluating Student Evaluations of Teaching: a Review of Measurement and Equity Bias in SETs and Recommendations for Ethical Reform." *Journal of Academic Ethics.* https://doi-org.southwesternu.idm.oclc.org/10.1007/s10805-021-09400-w

3. *Optional:* I'm also sharing a bibliography assembled for participants in the ACS Faculty of Color Uniting for Success program, which includes a variety of first-hand accounts of faculty experiences of student (and colleague) bias.

4. *Optional:* On implicit bias, generally: see the UCLA "Implicit Bias Video Series" from their "Faculty Search Committee Resources" guide for a brief overview of what research has shown about implicit bias.

Guidance for Faculty Evaluators in Using Student Evaluation Data

A few heads up here, too:

- You're going to be disappointed that there isn't more specific guidance about exactly *how* to bracket biases when reading student evaluations. What Michigan's CRLT says is simply to "pay attention to contextual factors that may influence results" when reviewing student evaluation data, for example. Linse (below) and others focus largely on best practices related to handling the quantitative results (though she does provide some useful guidance about interpreting qualitative data.)
- I do think it could be helpful to develop a list of the contextual factors that can influence student ratings -- for example, large class sizes, whether the course is required or elective, instructor gender, instructor race, etc. -- and keep an eye both on the individual items on the list and also on situations where they intersect -- where a course might have multiple biases at play (for example, a required lower-division course with female instructor of color).

1. University of Michigan Center for Research on Learning and Teaching, "Best Practices for Using Student Ratings for Personnel Decisions." (short)

2. Linse, Angela. (2017). "Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees," *Studies in Educational Evaluation* (54): 94-106.
    - See especially section 6, "Guidelines for faculty who use student ratings data to evaluate other faculty" (100-103).
    - Linse is considered one of the go-to sources on interpreting student ratings, but she's not explicitly addressing how to manage biases that may shape evaluations.

3. Esarey, Justin and Natalie Valdes. (2020). "Unbiased, Reliable, and Valid Student Evaluations Can Still Be Unfair." *Assessment & Evaluation in Higher Education* 45:8. DOI: 10.1080/02602938.2020.1724875
    - This article focuses on using quantitative measures for direct comparison of faculty -- a very common approach at many institutions -- and shows why it isn't empirically supportable.
    - *Abstract:* "Scholarly debate about Student Evaluations of Teaching (SETs) often focuses on whether SETs are valid, reliable, and unbiased. In this paper, we assume the most optimistic conditions for SETs that are supported by the empirical literature. Specifically, we assume that SETs are moderately correlated with teaching quality (student learning and instructional best practices), highly reliable, and do not systematically discriminate on any instructionally irrelevant basis. We use computational simulation to show that, under ideal circumstances, even careful and judicious use of SETs to assess faculty can produce an unacceptably high error rate: (a) a large difference in SET scores fails to reliably identify the best teacher in a pairwise comparison, and (b) more than a quarter of faculty with evaluations at or below the 20th percentile are above the median in

instructional quality. These problems are attributable to imprecision in the relationship between SETs and instructor quality that exists even when they are moderately correlated. Our simulation indicates that evaluating instruction using multiple imperfect measures, including but not limited to SETs, can produce a fairer and more useful result compared to using SETs alone."