EVALUATING STUDENT EVALUATIONS OF TEACHING

MEASUREMENT AND EQUITY BIAS IN SETS AND RECOMMENDATIONS FOR REFORM

Rebecca J. Kreitzer (UNC-Chapel Hill) and Jennie Sweet-Cushman (Chatham University)

2/15/2021

Standard Evaluations of Teaching (SETs) are commonly used in critical personnel decisions in academia. Despite their ubiquity, SETs are problematic because of their known measurement and equity biases. This brief summarizes the literature on SETs and makes recommendations on how to make SETs less problematic.

MEASUREMENT BIAS

Measurement bias occurs where variables unrelated to teaching effectiveness systematically influence the results in SETs. Common examples of course characteristics that cause measurement bias include class time, class size, if the class is a required or elective class, course difficulty, and discipline. Common examples of individual characteristics of students that cause measurement bias include the student's interest in the course material and their previous coursework.

Some of the ways in which evaluations vary:

- Classes with lighter workloads or higher grading distributions have higher scores.[i]
- Evaluations are lower for non-elective and quantitative courses.[ii]
- Evaluations are higher for upper-level, discussion-based classes compared to larger, introductory courses.[iii]
- Evaluations vary across disciplines; natural science courses receive the lowest scores and humanities the highest.[iv]
- Bringing cookies or chocolate to class increases evaluation scores.[v]

> "[T]he 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."[vi]
>
> 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."[vii]

In summary, evaluations are shaped by course and individual characteristics unrelated to actual instructor quality. On this basis alone, universities and colleges should reconsider the use of these evaluations in high stakes employment decisions, such as hiring and for promotion.

EQUITY BIAS

Equity bias occurs when variables outside the instructor's control systematically influence the results. Common examples of instructor characteristics that cause equity bias include instructor's gender, race, ethnicity, accent, sexual orientation or disability. The evidence of equity bias is strongest in the qualitative comments about the course or instructor.

Evidence of gender equity bias includes:

- Male instructors are perceived as more accurate in their teaching, have more education, are less sexist, more enthusiastic, competent, organized, professional, effective, easier to understand, prompt in providing feedback, and are less penalized for being tough graders.[viii]
- Experimental designs that manipulate the gender of online instructors find that instructors receive lower evaluations when students believe their instructor is a woman, despite identical course delivery.[ix]
- Women receive lower scores in the social sciences and higher scores in the humanities.[x] There is no discipline where women receive higher evaluative scores compared to men in online evaluations.[xi]
- There is a gender affinity effect, whereby students prefer instructors of the same gender.[xii] This affinity is also likely the case with race, though there is not currently any research on this.

Conforming to prescribed gender roles has a more significant effect than gender itself.[xiii] Women faculty are rated highly for exhibiting traditionally feminine traits, like warmth and sensitivity, by male and female students.[xiv] Men are evaluated

positively on traditionally male traits, like perceptions of intelligence.[xv] These gender stereotypes harm women instructors because students prefer professors with masculine traits, yet penalize women for not conforming to feminine stereotypes.[xvi] Because of the conditional nature of equity bias, there are not many estimates of the size of its effect. One study found that, controlling for other factors, female instructors received average ratings .5 standard deviation lower than male instructors' ratings.[xvii]

There is far less research on equity bias in teaching evaluations for faculty of color, in part because of their severe underrepresentation in academia. Faculty of color are evaluated worse than their male colleagues, especially Black and Asian professors, with Black male professors fairly particularly poorly.[xviii] Faculty with accents and Asian last names fare worse than their native English-speaking counterparts.[xix] People of color are also punished for not conforming to intersectional stereotypes.[xx]

There is some evidence to suggest that LGBTQ faculty fare worse than their straight colleagues.[xxi] Some research indicates seniority decreases bias,[xxii] while other research finds that younger professors are more popular and receive higher evaluations.[xxiii] We know almost nothing about biases against other relevant intersectional identities, such as pregnancy or disability.[xxiv]

RECOMMENDATIONS FOR BETTER USE OF STUDENT EVALUATIONS

1. Contextualize evaluations of students' experiences, not a measure of teaching.[xxv] Students cannot, and arguably should not, evaluate teaching. When contextualized as surveys of students' perceptions and experiences, they can provide useful feedback for faculty and administrators.

2. Be proactive about increasing the validity of these assessments by improving the response rate. A lower response rate is more likely to be unrepresentative.[xxvi]

3. Administrators should interpret the results of student ratings with caution. Student evaluations were not designed to be used as a comparative metric across faculty.[xxvii] Evaluations should be used to compare a faculty member's trajectory of teaching over time, and ideally, within a single course.[xxviii] Because the distribution of most faculty members' reviews have a negative skew,[xxix] administrators should look at the median or modal response, rather than the mean. Mean the distribution of evaluations is not normally distributed, means may be biased.

4. Restrict or eliminate the use of qualitative comments, which have the strongest evidence of equity bias. Women faculty and faculty of color are more likely to receive negative comments about personality traits, appearance, mannerisms, competence, and professionalism.[xxx] Instead of asking for general "comments," assessments should direct students to provide feedback in response to specific prompts.

   Qualitative comments are problematic for many reasons, including being difficult to aggregate because of small sample sizes[xxxi]; they are often contradictory and not reliable;[xxxii] suffer from novelty bias (people are more likely to remember unexpected or uncommon comments) and negativity bias (people are more likely to remember negative information).

5. Administrators should not rely on SETs as the sole method of assessing teaching. There are several alternatives or supplements to SETs, including: peer observation[xxxiii], comprehensive evaluations of teaching portfolios,[xxxiv] and reviews of course materials.[xxxv] While these alternatives may also be susceptible to biases, they are not systematically bias in the same way.[xxxvi] Several imperfect measures are better than using just one.

6. There should be more research on interventions to reduce bias. There are only a few articles on testing interventions to reduce equity bias. Reducing the size of the scale can mitigate gender bias.[xxxvii] One random control trial (RCT) finds that making students aware of biases can mitigate the gender gap in SETs,[xxxviii] while another RCT finds the opposite.[xxxix] Anecdotal evidence suggests there may be a backlash effect when under-represented groups discuss equity bias in SETs with students.

This policy brief provides a summary of a peer-reviewed article.
Suggested citation: Kreitzer, R. J., & Sweet-Cushman, J. (2021). Evaluating Student Evaluations of Teaching: a Review of Measurement and Equity Bias in SETs and Recommendations for Ethical Reform. *Journal of Academic Ethics*, 1-12.

[i] Greenwald & Gillmore, 1997; Miles & House, 2015; Rosen, 2018; Sinclair & Kunda, 2000
[ii] Benton & Cashin, 2012; Boring et al, 2016; Chamberlin & Hickey, 2001; Elmore & LaPointe, 1975; Greenwald & Gillmore, 1997; Mengel et al 2018; Uttel et al, 2017, 2013
[iii] Hamermesh & Parker 2005; Miles & House, 2015; Sidanius & Crane, 1989; Sporen et al, 2013; Centra & Gaubatz, 1998
[iv] Basow & Montgomery, 2005; Basow & Silberg, 1987
[v] Hessler et al, 2018; Youmans & Jee, 2007
[vi] Uttel et al, 2017, 19
[vii] Esarey & Valdes, 2020, 1
[viii] Abel & Meltzer, 2007; Arbuckle & Williams, 2004; 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
[ix] Boring et al, 2016; MacNell et al 2015
[x] Basow & Montgomery, 2005
[xi] Rosen, 2018
[xii] Bachen et al 1999, Martin, 1984, Young et al, 2009; Basow & Silberg, 1987; Basow, 1995; Burns-Glover & Veith, 1995; Fan et al, 2019; Kaschak, 1981, 1978; Mengel et al, 2018; Bray & Howard, 1980; Centra, 2000, Rowden & Carlson, 1996
[xiii] Basow & Silberg, 1987; Freeman, 1994
[xiv] Wallisch & Karau, 2002
[xv] Bian et al, 2017; Basow, 2000; Boring, 2017; Leslie et al, 2015
[xvi] Burns-Glover & Veith, 1995; Bennett, 1982; Boring, 2017; Kierstead et al, 1988
[xvii] Hamermesh & Parker, 2005
[xviii] Reid 2010; Chavez & Mitchell, 2020
[xix] Fan et al, 2019; Subtirelu, 2015
[xx] Anderson, 2010, Anderson & Smith, 2005
[xxi] Anderson & Kannerm 2011; Ewing et al, 2013
[xxii] Mengel et al, 2018; Wiginton et al, 1989
[xxiii] Arbuckle & Williams, 2003; McPherson et al, 2009
[xxiv] But see, Baker & Copp, 1997
[xxv] Linse, 2017; Abrami, 2001; Arreola, 2004
[xxvi] Chapman & Joines, 2017; Adams & Umbach 2012
[xxvii] Franklin 2001
[xxviii] Linse, 2017
[xxix] Linse, 2017; Arreola, 2004; Hativa 2013a, b
[xxx] Wallace et al, 2019
[xxxi] Himelein, 2018
[xxxii] Linse, 2017
[xxxiii] Miller & SHeldin, 2014
[xxxiv] Centra, 2000; Seldin et al 2010
[xxxv] Chism, 2007
[xxxvi] Esarey & Valdes, 2020
[xxxvii] Rivera & Tilcsik, 2019
[xxxviii] Peterson et al, 2019
[xxxix] Key and Ardoin, 2019