Even ‘Valid’ Student Evaluations Are ‘Unfair'

insidehighered.com/news/2020/02/27/study-student-evaluations-teaching-are-deeply-flawed

Student evaluations of teaching reflect students’ biases and are otherwise unreliable. So goes much of criticism of these evaluations, or SETs. Increasingly, research backs up both of those concerns.

On the other side of the debate, SET proponents acknowledge that these evaluations are imperfect indicators of teaching quality. Still, proponents argue that well-designed SETs inevitably tell us something valuable about students’ learning experiences with a given professor.

A new study -- which one expert called a possible “game-changer” -- seeks to cut through the noise by assuming the best of SETs -- at least, that which is supported by the existing literature. Its analysis assumes that the scores students give instructors are moderately correlated with student learning and the use of pedagogical best practices. It assumes that SETs are highly reliable, or that professors consistently get the same ratings. And it assumes that SETs do not systematically discriminate against instructors on the basis of irrelevant criteria such as their gender, class size and type of course being taught.

And even when stacking the deck for SETs, the study finds that these evaluations are deeply flawed measures of teaching quality.

New Question, Familiar Answer

"Unbiased, Reliable and Valid Student Evaluations Can Still Be Unfair," published in *Assessment & Evaluation in Higher Education*, was written by Justin Esarey and Natalie Valdes. Esarey, an associate professor, and Valdes, an undergraduate research fellow, both work in political science at Wake Forest University. They note -- rightly -- that their field has faced concerns about gender bias, including in student evaluations of female professors.

The problem transcends political science, of course, and many studies suggest that students perceive instructors differently based on factors beyond gender, such as race. (Political scientists Mirya Hollman, Ellen Key and Rebecca Kreitzer maintain a bibliography of relevant studies here.)

As the paper notes, "Using invalid, unreliable or biased student evaluations to make decisions about hiring and tenure is obviously harmful to students and faculty alike." Even worse, it says, "biased SETs could disadvantage faculty from underrepresented minority groups or punish faculty members who teach unpopular required courses."

While these are "important problems," the authors write, they shift gears and "ask a different question: if SETs are valid, reliable, and unbiased, what then?" Are SET scores without "demonstrable bias and moderately correlated with instructor quality a fair basis on which to judge a faculty member's teaching performance?" If the answer to the latter question is no, then "there is a much bigger problem with the use of SETs than is commonly recognized."

And no is indeed the answer: even under "ideal" circumstances, Esarey and Valdes write, SETs still yield an "unacceptably high error rate."

Summing up his findings this week, Esarey said that unless the correlation between student ratings and teaching quality is "far, far stronger than even the most optimistic empirical research can support," then common administrative uses of SETs "very frequently lead to incorrect decisions." Those professors with the very highest evaluations "are often poor teachers," he added, "and those with the very lowest evaluations are often better than the typical instructor."

Consequently, Esarey said that he and Valdes would expect "any administrative decisions made using SET scores as the primary basis for judgment to be quite unfair."

Experts in this area have long advised against basing high-stakes personnel decisions on student ratings of instruction alone. A number of institutions and professional groups have made commitments and policy changes to this effect. But SETs still have a major foothold in

these processes on many campuses, as they are relatively easy and inexpensive compared to other means of assessing teaching quality. And because institutions invest relatively little time and few resources in their adjunct faculty members, these professors are disproportionately hired and fired based on student feedback.

Benefit of the Doubt

The current study is based on a computational simulation -- no actual professors were involved (or harmed). That allowed Esarey and Valdes to directly measure teaching effectiveness, which is still very hard to measure in real life. For the same reason, Esarey and Valdes were also able to assess how accurate are administrative decisions using SET scores to gauge teaching effectiveness.

As Esarey explained, “In our simulation, we know a faculty member's SET score and also their real teaching effectiveness. We computationally simulate thousands of faculty members and then compare them to one another the way that a department chair or dean might evaluate faculty members using SET scores in real life.”

A bit more technically, the complex computer model simulated one million instructors' student ratings along with their teaching quality percentiles, with varying correlation between the two measures. Then it used the simulated scores in realistic evaluation scenarios. First, Esarey and Valdes looked at “pairwise comparisons” of sets of hypothetical faculty members via SET scores. This mirrored “comparison of job candidates on the basis of their teaching performance or the comparison of a faculty member up for tenure to the teaching record of a recent (un)successful case,” according to the study.

Next, Esarey and Valdes compared an individual professor’s SET scores to the overall population of SET scores from all faculty members in the model. That, in turn, mirrored a procedure “where faculty members who are under-performing relative to their peers (e.g. whose scores are below a certain percentile ranking) are identified for administrative action as part of a tenure case or other systematic review,” the study says.

In so doing, the researchers found that even when the correlation between instructor ratings and faculty instructional quality or student learning is as significant as it's ever been shown to be (about 0.43, based on a 1981 metastudy that has since been challenged), there remains a large difference in SET scores -- as much as 30 percentage points. This does not reliably identify the best teacher in the pairwise comparison.

Moreover, one-quarter of these simulated faculty members with SET scores at or below the 20th percentile in the peer comparison analysis “are actually better at teaching than the median faculty member in our simulation.”

Even those with exceptionally high SET scores can be “poor teachers,” the study says, as nearly 19 percent of those with SET scores above the 95th percentile are no better than the

median professor at teaching.

Making “fair, accurate personnel decisions based on faculty instruction requires a measure of teaching performance that is substantially more related to student learning or instructional best practices than SET scores alone,” the study says. (The researchers confirmed their findings in a second, more advanced analysis.)

As for how SETs should be used within colleges and universities, the researchers make three recommendations. On a technical level, they advise removing any systematic gap in SET scores explained by noninstructional factors, such as gender, via regression adjustment or matched subsample analysis “before using these scores for any purpose.”

How to Use SETs

This kind of adjustment can’t “filter” out all idiosyncratic influences on SET scores, however, they say. They thus advise using a “combination of independent evaluators, interviews with students, teaching observations by experts, peer review of instructional materials and SET scores” to give “a much more accurate picture of a faculty member’s teaching proficiency when SET scores alone would be misleading.”

Averaging these multiple forms of evaluation can allow idiosyncratic variation in each one to cancel out, “resulting in further reduction of imprecision between the averaged assessment and a faculty member’s true teaching performance,” the study says.

Because this kind of multifaceted assessment is expensive, the researchers say that SETs “could serve as a low-cost mechanism for identifying” professors who need it -- but only “with the understanding that many faculty so identified will be excellent teachers.”

Last, the authors advise “caution in over-reliance on SET scores for any purpose.”

Joshua Eyler, director of faculty development at the University of Mississippi and author of *How Humans Learn: The Science and Stories Behind Effective College Teaching*, commented on a study draft prior to publication. Evidently pleased with the results, he’s the one who called the study a “game-changer” in the SET wars.

Eyler said this week that there is a big difference between asking students about a professor's "behaviors" -- whether they have a sense of humor or they're engaging -- and observing whether professors are using evidenced-based teaching strategies. That's because behaviors are rarely if ever correlated with student learning, whereas good strategies are.

With regard to SETs in particular, Eyler said that if an institution uses a form that poses real questions linked to student learning (and not behaviors), then SETS “have a role to play in providing formative, nonevaluative feedback for faculty.” Yet they “should simply not be

used for summative evaluations and decisions about someone's career," he cautioned, as the study makes clear that "even in a perfect world where we could somehow mitigate the bias of SETs, they would still be deeply flawed instruments."

Esarey said he endorsed what he called "multi-modal" assessments of teaching. Echoing him, Eyler said that the best tenure and promotion practices "employ multiple modes of evidence for teaching effectiveness."