Bias Against Female Instructors

insidehighered.com/news/2016/01/11/new-analysis-offers-more-evidence-against-student-evaluations-teaching

There's mounting evidence suggesting that student evaluations of teaching are unreliable. But are these evaluations, commonly referred to as SET, so bad that they're actually better at gauging students' gender bias and grade expectations than they are at measuring teaching effectiveness? A new paper argues that's the case, and that evaluations are biased against female instructors in particular in so many ways that adjusting them for that bias is impossible.

Moreover, the paper says, gender biases about instructors -- which vary by discipline, student gender and other factors -- affect how students rate even supposedly objective practices, such as how quickly assignments are graded. And these biases can be large enough to cause more effective instructors to get lower teaching ratings than instructors who prove less effective by other measures, according to the study based on analyses of data sets from one French and one U.S. institution.

"In two very different universities and in a broad range of course topics, SET measure students' gender biases better than they measure the instructor's teaching effectiveness," the paper says. "Overall, SET disadvantage female instructors. There is no evidence that this is the exception rather than the rule."

Accordingly, the "onus should be on universities that rely on SET for employment decisions to provide convincing affirmative evidence that such reliance does not have disparate impact on women, underrepresented minorities, or other protected groups," the paper says. Absent such specific evidence, "SET should not be used for personnel decisions."

"Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness," was published last week in *ScienceOpen Research*. Philip B. Stark, associate dean of the Division of Mathematical and Physical Sciences and a professor of statistics at the University of California at Berkeley and co-author of a widely read 2014 paper questioning the reliability of evaluations, co-wrote the paper with Anne Boring, a postdoctoral researcher in economics at the Paris Institute of Political Studies, and Kellie Ottoboni, a Ph.D. candidate in statistics at Berkeley.

For their study, Stark and his colleagues performed advanced statistical analyses of five years' worth of data to which Boring had access regarding 23,001 evaluations of 379 instructors by 4,423 students in six mandatory first-year courses at a French university. They also applied the tests to evaluations for four sections of an online course in a randomized, controlled, blind experiment at a U.S. university that was the data set for another popular 2014 paper on gender bias in student teaching evaluations. (In that study, co-written by Lillian MacNell, a Ph.D. candidate in the department of sociology and anthropology at North Carolina State University at Raleigh, students in an online course on technology and society gave better evaluations to the teaching assistants they thought were male, even when the two instructors -- one male and one female -- had switched their identities. They both used both identities, which made it possible to compare what happened when each was apparently female.)

The idea for the new study was to investigate whether student evaluations of teaching primarily measure teaching effectiveness or biases using a higher level of statistical rigor than had previously been applied to the data sets. Their method was to use nonparametric permutation tests, statistical tests of significance for hypotheses such as "any given student would rate two instructors the same if the instructors are identical except for their apparent gender."

Through these tests, Stark and his co-authors found that the association between evaluations and a more objective measure of teaching effectiveness -- student performance on an anonymously graded final in the French data set (grades were not linked to the evaluations in the U.S. set) -- is weak, and not statistically significant. Yet the association between evaluations and perceived instructor gender in both the U.S. and French data sets is largely

statistically significant: instructors whom students believe are male receive significantly higher average ratings.

Different Biases, Same Outcome

Students' gender appeared to impact their bias, but in different ways in the French and U.S. samples.

In the French data, male students tended to rate male instructors higher than they rated female instructors, but little difference was observed among female students. In the U.S. data, female students tended to rate perceived male instructors higher than they rated perceived female instructors, with little difference in ratings by male students. In both cases, however, the bias still positively impacted male instructors and disadvantaged female ones.

Stark said in an email interview that this difference -- leading to the same outcome -- was the most surprising finding of the study. At one university, he said, "male students rate male instructors higher, although they apparently learn less from male instructors. In the other, female students rate (apparently) male instructors higher."

The paper considers whether men receive better overall scores because they're better instructors, by analyzing the relationship between instructor gender and students' average final exam score. In all disciplines in the French sample, students of male instructors performed worse -- though not in ways that are statistically significant.

So why do male instructors receive higher scores? A separate analysis by student gender in the French data suggest that male students give higher scores to male instructors, especially in history (p-value of 0.01), microeconomics (p-value of 0.01), macroeconomics (p-value of 0.04) and political science and political institutions (with p-values of 0.06 and 0.08, respectively). The effect was not statistically significant in sociology (p-value of 0.16). (Smaller p values generally indicate evidence against the null hypothesis, or the assumption being tested.)

"The average correlation between instructor gender and SET is statistically significant -- male instructors get higher SET -- but if anything, students of male instructors do worse on final exams than students of female instructors," the paper says. "Male students tend to give male instructors higher SET, even though they might be learning less than they do from female instructors. We conclude that SET are influenced more by instructor gender and student grade expectations than by teaching effectiveness."

Controlling for Teaching Styles

The French "natural experiment," which happened naturally over five years, didn't allow the researchers to control for differences in teaching styles, but the U.S. data (MacNell and her collaborators' data from North Carolina State) did. That experiment collected evaluation data from an online course in which 43 students were randomly assigned to four discussion groups, each taught by one of two teaching assistants -- one male and one female. In one group taught by the male instructor, he used the female instructor's name, and vice versa. The instructors gave similar feedback to students and returned assignments at exactly the same time.

MacNell in 2014 found that "the male identity received significantly higher scores on professionalism, promptness, fairness, respectfulness, enthusiasm, giving praise and the student ratings index. … Students in the two groups that perceived their assistant instructor to be male rated their instructor significantly higher than did the students in the two groups that perceived their assistant instructor to be female."

Whereas MacNell used parametric tests whose assumptions did not meet her study's experimental design, Stark and his co-authors used permutation tests that are consistent with the underlying randomization and avoid parametric assumptions about evaluations. They also looked at some new effects, such as the interaction between student gender and perceived instructor gender.

They say the new analysis supports MacNell and her colleagues' overall conclusions, and in some instances -- such as bias regarding promptness -- more strongly. (Although in other cases, such as knowledgeability, the new analysis found evidence for the effect to be weaker than originally suggested.) Since assignments were returned at exactly

the same time in all four sections, the significantly lower rating for female instructors (what equates to about 16 percent of full scale) “seriously impugns the ability of SET to measure even putatively objective characteristics of teaching,” the paper reads.

Again, Stark and his colleagues found that, in contrast to the French data, perceived male instructors were rated significantly more highly not by male students but by female students. Male students rated the perceived male instructor somewhat significantly higher on only one criterion -- fairness (p-value 0.09). But female students in the U.S. sample rated the perceived male instructor higher on overall satisfaction (p-value 0.11) and most aspects of teaching. Those include praise (p-value 0.01), enthusiasm (p-value 0.05) and fairness (p-value 0.04).

Female students rated perceived female instructors lower on helpfulness, promptness, consistency, responsiveness, knowledge and clarity, although the differences are not statistically significant, the paper says.

MacNell, the author of the original paper concerning the U.S. data, said she’d seen the new study and agreed that student evaluations of teaching “generally do not accurately measure teaching effectiveness.”

Since this holds true across a variety of settings -- different courses, institutions and departments -- “there may not be much we can do to address this fact,” she said. At the same time, she added, these evaluations can still be useful tools in limited ways.

“Perhaps institutions can move away from using SET for decisions about hiring, promotion and tenure,” she said, “but still use them to get feedback on what students want and expect from their courses.”

Stark said he doubted the new study would be the “nail the coffin” for student evaluations of teaching, but said he hoped it will “bring us closer to ending any use of SET for employment decisions.” Still, he said, pretending that such evaluations are strong measures of teaching effectiveness remains “irresistible” to some, for a variety of complicated reasons.

But could the tide be turning? Stark said he expected class action lawsuits against universities that rely on these evaluations for employment decisions will start this year, and that there’s evidence to support such cases.

“Our analysis would support an argument that the use of SET has adverse impact on female instructors, at least in the two settings we examined,” he said. “Replication of this kind of experiment and analysis elsewhere would strengthen the argument. Eventually, lawsuits will lead universities to do the right thing, if only to mitigate financial risks.”