Zero Correlation Between Evaluations and Learning

insidehighered.com/news/2016/09/21/new-study-could-be-another-nail-coffin-validity-student-evaluations-teaching

A number of studies suggest that student evaluations of teaching are unreliable due to various kinds of biases against instructors. (Here's one addressing gender.) Yet conventional wisdom remains that students learn best from highly rated instructors; tenure cases have even hinged on it.

What if the data backing up conventional wisdom were off? A new study suggests that past analyses linking student achievement to high student teaching evaluation ratings are flawed, a mere "artifact of small sample sized studies and publication bias."

"Whereas the small sample sized studies showed large and moderate correlation, the large sample sized studies showed no or only minimal correlation between [student evaluations of teaching, or SET] ratings and learning," reads the study, in press with *Studies in Educational Evaluation*. "Our up-to-date meta-analysis of all multisection studies revealed no significant correlations between [evaluation] ratings and learning."

These findings "suggest that institutions focused on student learning and career success may want to abandon SET ratings as a measure of faculty's teaching effectiveness," the study says.

The paper considered end-of-course evaluations, not arguably more subjective ratings found on ratings websites.

Authors of the new paper scrutinized data taken from seven studies that have been cited over time as evidence of the effectiveness of student evaluations. Some of the data, for example, come from a 1981 meta-analysis of multisection validity studies. That analysis, based on 41 studies reporting on 68 multisection courses, found a significant link between overall instructor course rating and student achievement, especially on measures of skill and structure. It endorsed student ratings as valid measures of teacher effectiveness.

Yet, according to the new analysis, that paper and others like it "suffer from multiple critical methodological flaws that render their conclusions unwarranted." Namely, the studies fail to do some or all of the following: provide basic information about the primary-level data, such as effect and sample size; ensure the data's accuracy, such as by checking how they're coded; or, perhaps most importantly, consider small sample size bias. The latter occurs when statistical results that may not be representative of the sample as a whole are gathered or reported in such a way that shows significant -- and therefore more likely to be published -- results.

The 1981 study, for example, did briefly consider sample size, in terms of course sections, but reported it was not a significant factor. A few pages later, the same study dismissed reviewers "concerned that rating/achievement correlations vary according to the number of sections used in the study," but then somewhat inexplicably said a "number of sections correlated significantly with the absolute value of effect size." Correlation size was not reported.

A rerunning of that study's original, available data found that the number of sections included in multisection studies was generally small, with the number of multisection studies based on as few as five sections, and that "many impossibly high correlations ($r > 0.90$) were obtained in multisection studies with a small number of sections." It also found that the majority of reported rating-achievement correlations were not statistically significant, and that the magnitude of evaluation-achievement correlations decreased for larger-sized studies in a predictable pattern.

The study says that the best evidence -- its own meta-analysis of SET-learning correlations when prior learning and ability are taken into account -- indicates that the SET-learning correlation is actually zero, and that it's "astonishing" that poor data have driven the conversation around evaluations for some 30 years. The paper advises universities to begin giving teaching evaluations appropriate "weight" in personnel and other decisions.

"The entire notion that we could measure professors' teaching effectiveness by simple ways such as asking

students to answer a few questions about their perceptions of their course experiences, instructors' knowledge and the like seems unrealistic given well-established findings from cognitive sciences such as strong associations between learning and individual differences including prior knowledge, intelligence, motivation and interest," the paper says. "Individual differences in knowledge and intelligence are likely to influence how much students learn in the same course taught by the same professor."

Small sample size bias concerns aren't unique to student evaluations of teaching -- it's a concern in neuroscience, for example, and many other fields. But the new analysis is one more reason for critics to question the validity of student evaluations of teaching as effective measures of quality. A recent Stanford University investigation of meta-analyses also found them to be problematic. "Currently, there is massive production of unnecessary, misleading, and conflicted systematic reviews and meta-analyses," that paper says. "Instead of promoting evidence-based medicine and health care, these instruments often serve mostly as easily produced publishable units or marketing tools."

"Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related" was written by Bob Uttl, professor of psychology at Mount Royal University; Carmela A. White, a graduate student in psychology at the University of British Columbia; and Daniela Wong Gonzalez, a graduate student at the University of Windsor, all in Canada. Most of the studies analyzed were based on U.S. data.

Philip B. Stark, associate dean of the Division of Mathematical and Physical Sciences and a professor of statistics at Stanford, is a vocal critic of teaching evaluations used as high-stakes measures of teaching effectiveness (he did not write the recent study on meta-analysis). He said Uttl's and his colleagues' paper "pays much more attention than usual to the quality of the underlying studies, and gives a circumspect review of previous meta-analyses."

Given what "the best randomized, controlled experiments have shown, it is not surprising that this study finds no meaningful correlation between SET and learning," he said. And given the "strong association between SET and instructor gender, this adds evidence to the argument that institutions that care about learning should abandon SET as a measure of teaching effectiveness."

Uttl said that contrary to popular belief, "the multisection studies do not support validity of SET ratings as measure of faculty's teaching effectiveness. They indicate that students do not learn more from professors with higher SET ratings."