Dror Bar-Natan: Academic Pensieve: 2022-05: http://drorbn.net/ap/2022-05/BewareOfSETs-Comments.pdf

Comments on “Beware of Teaching Evaluations!” by Alfonso Gracia-Saz. My comments are in blue; the original document is in red and black, with some hyperlinks in light blue.

Student Evaluations of Teaching (SETs) are not a good way to measure teaching effectiveness or student learning.

Neither are “publications” (for research), nor are “letters of recommendations” (for anything). Let us not delude ourselves that we know how to measure anything well.

Clarification: my SET scores have always been above average. What I write here is not an attempt to explain away my bad ratings because I have never had bad ratings. However, precisely because of that, I take it as my moral obligation to raise this issue on behalf of colleagues who would not be taken seriously if they raised it themselves.

My teaching evaluations are also usually reasonably good. This masks my deep inconfidence. I never know if I really do well, and much of the time I think that I don’t.

And my own clarification: Much as I love my own research, from a societal perspective I believe that the justification for the existence of university mathematics is teaching, not research. Research is mostly there to make sure that the teachers have excellent knowledge, the kind that gets lost if one does not actively live what they preach. Research may be a bit of a peacock’s tail: it might have evolved to be a bit too big.

There is ample evidence (see below) that whatever SETs measure, it is not how much students are learning, and that good SET ratings do not mean good teaching. Over-reliance on SETs lowers the quality of our teaching, stifles innovation, penalizes instructors who take risks, and encourages giving easy As rather than pushing students to realize their full potential. Having been in various hiring and evaluation committees I am cynical about the use of SETs. More often than not, I have seen SET ratings as the main or even the only criterion to evaluate instructors, contrary to what those same committees might have claimed. Professors used to teaching only small courses, or advanced courses, or courses for specialists, and who never deal with large, first-year, service courses are particularly likely to fall into this trap.

I agree that all these traps that Alfonso mentions are easy to fall into and are serious concerns.

You don't need to take my word for it. Look at the research.

The USAFA Study

The "US Air Force Academy Study" by Carrell and West (2010) was done under some ideal circumstances that few other studies in education normally achieve. As a summary:

- More than 10,000 students were involved in the study over a period of 7 years.
- All students, regardless of major, were required to take the same core sequences of courses, including Calculus I and various courses that had Calculus I as a prerequisite. The researchers looked at their performance in the follow-up courses as an unbiased measure of the quality of their learning of Calculus I.
- Average class size was 20 students. Students were assigned to different instructors randomly.
- All instructors shared the same syllabus and gave the same final exam. They graded the final exam collectively ("Professor A grades Question 1 for all students") to maintain homogeneous standards.

If we define teaching effectiveness as how well student do in the next course, then **SET scores correlate *negatively* with teaching effectiveness**. More specifically, Carrell and West found that students who give good SET scores in Calculus I perform better in Calculus I but perform worse in follow-up courses. This correlation applied to every question in the SET questionary individually.

The statement in red is a sensational yet false deduction from the “more specifically” explanation which follows it. The effect that Carrell and West see may well be a standard “regression to the mean” effect: Students who love their profs do better in that specific class and then regress to their normal level of achievement. (There may also be a small “better students are more critical” effect; who knows the specifics of USAFA?). The relevant comparison is between the achievements of graduates of different sections. If this comparison was made, it is not presented here.

Carrell and West offer an explanation on which I will elaborate further. An instructor has to decide whether to spend time "preparing students for the test" or creating what they call "deep learning".

- An instructor who focuses on preparing students for the test will be rewarded with good SET scores. Her students will perform well on the final exam, but the final exam will no longer be an accurate measure of learning. Indeed, based on their weaker performance on follow-up courses, her students have not learned much.
- An instructor who focuses on deep learning has added real, long-lasting value. Her students will resent this, since it requires more work, and will penalize her with lower SET scores. But they have learned more, as evidence by their better performance in follow-up courses.

This is the Carrell and West explanation. There is a component of truth in it, I'm sure. But the evidence presented above also has an alternative explanation, as I have already indicated.

BTW, from a student's perspective, the Carrell and West explanation is patronizing and insulting. They essentially tell the students that they don't know what's good for them. I'm glad this is not the whole truth.

Moreover, the researchers found that the more experienced instructors were more likely to have lower SET scores (and better student performance in follow-up courses, of course).

That may be ageism. It sucks, but it's a different story.

Other Findings

- A different randomized study also finds that SET scores are negatively correlated with student performance in follow-up courses. (Braga, Paccagnella, and Pellizzari, 2014)
- SET scores are not correlated with student learning. (Uttl, White, and Wong Gonzalez, 2017)
- SET scores are correlated with prior subject interest (Marsh and Cooper, 1981) and with grade expectations (Worthington, 2002).

Not surprising.

- SET scores can be predicted from students reaction to watching 30 seconds of silent video of their instructor before the course. (Ambady and Rosenthal, 1993)

"Predicted", or "correlated with"? "Correlated with" is not surprising and is at least partially legitimate: A teacher with their back to the students or with obvious lack of enthusiasm is a worse teacher, and a 30 second video may be enough to pick that out.

- Students give lower SET scores to teaching innovation, even when it is good for them. (Smith and Cardaciotto, 2011)

This really is sad, if it is supported by the evidence.

- For a more thorough review of the research literature on SETs, and for recommendations, see (Stark and Freihstat, 2014) and (Hornstein, 2017).

My personal conclusion is that teaching evaluations are a limited tool, but not an empty tool. Perhaps the distinction between a 3.7 and a 4.3 is almost always too minor to matter, and when bigger differences arise, we should be aware of the biases that might have created them. Yet not penalizing a 2.0 instructor (barring truly unusual circumstances) is a mistake, and an insult to our students. Yes, I know, we should read the comments and not just the numerical values. The comments for a 2.0 instructor will nearly always be "they are the worst I've ever seen, and I haven't learned anything". Even if these students do actually learn something, they will come out of their course hating mathematics and hating our department. We don't want that.

I'm not sure what better tool we have. Other tools have other biases: LORs written before breakfast are different from LORs written after breakfast, self-reporting correlates with one's ego more than with anything else, and word of mouth is word of mouth.

References

1. N. Ambady and R. Rosenthal: *Half a minute: predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness*, Journal of Personality and Social Psychology, 64(3), 431-441, 1993, (pdf copy)
2. M. Braga, M. Paccagnella, and M. Pellizzari: *Evaluating students' evaluations of professors*, Economics of Education Review, 41, 71-88, 2014. (pdf copy)
3. S.E. Carrell and J.E. West: *Does professor quality matter? Evidence from random assignment of students to professors*, Journal of Political Economy, 118(3), 409-432, 2010. (pdf copy)
4. H.A. Hornstein: *Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance*, Cogent Education, 4(1), 2017. (pdf copy)
5. H.W. Marsh and T.L. Cooper: *Prior subject interest, students' evaluations, snd instructional effectiveness*, Multivariate Behavioral Research, 16(1), 83-104, 1981. (link)
6. C.V. Smith and L. Cardaciotto: *Is active learning like broccoli? Student perceptions of active learning in large lecture classes*, Journal of the Scholarship of Teaching and Learning, 11(1), 53-61, 2011, (pdf copy)
7. P.B. Stark and R. Freishtat: *An evaluation of course evaluations*, ScienceOpenResearch, DOI: 10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1, 2014. (pdf copy)
8. B. Uttl, C.A. White, and D.Wong Gonzalez: *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, 2017. (link)
9. A.C. Worthington: *The impact of student perceptions and characteristics on teaching evaluations: a case study in finance education*, Assessment and evaluation in higher education, 27(1), 49-64, 2002. (pdf copy)