Excellence or ease? Exploring student evaluations of teaching

Stroebe, Wolfgang

Published in:
Psychologist

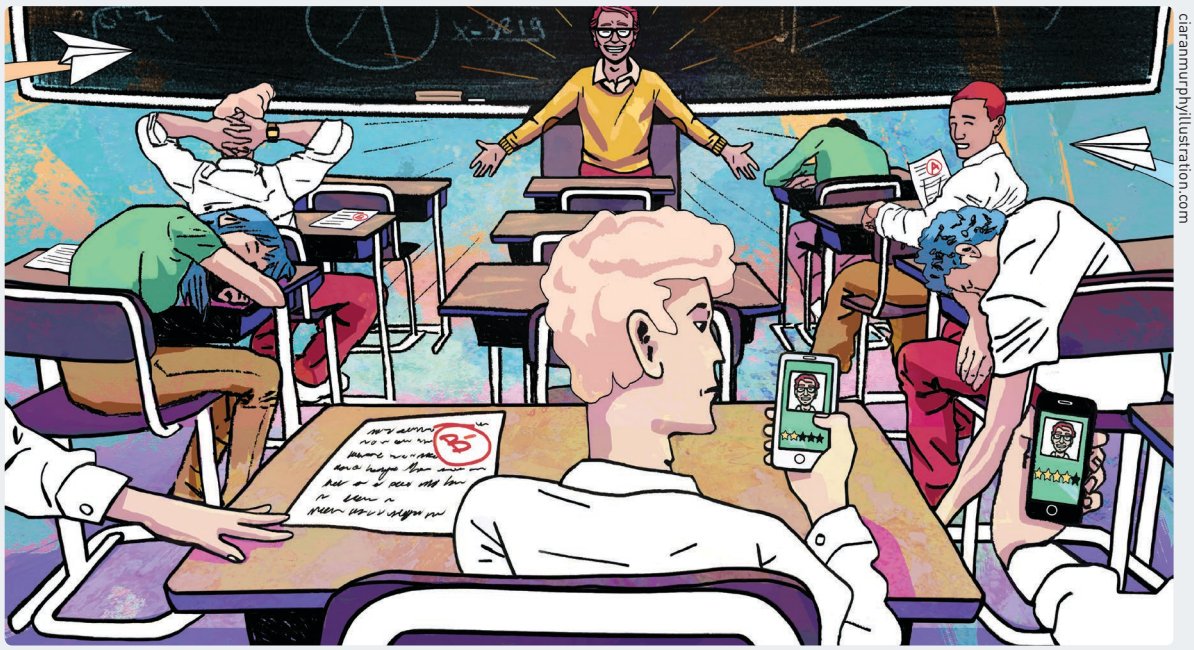
Excellence or ease? Exploring student evaluations of teaching

Wolfgang Stroebe is not a fan of student satisfaction measures...

When Hermann Remmers and Edwin Guthrie proposed Student Evaluation of Teaching (SETs) questionnaires in 1927, their sole intention was to help instructors to improve their teaching. Little did they know the measures might ultimately contribute to grade inflation and a decrease in course quality.

Student Evaluation of Teaching questionnaires are now widely used in many countries. In the United States, department chairs and deans use SETs as 'objective' information about teaching effectiveness in decisions about tenure and merit increases (Stroebe, 2016). Although this is (not yet) a practice in Britain, student satisfaction ratings are used as one of the indicators of teaching quality at British universities by the Teaching Excellence Framework.

Administrators find SETs very useful, because these questionnaires have a great deal of face validity. After all, student answers on how they rate their instructors, how informative they found their classes, how they would rate the courses overall, should provide important information about the teaching effectiveness of the instructor and about student learning in a



ciaranmurphyillustration.com

course. The problem is that – despite appearances – SETs are invalid as measures of teaching effectiveness and student learning. In fact, there is some indication that it is sometimes the bad teachers who receive good evaluations.

The validity of SETs

SETs would not have become as popular as they are, if their validity had not been assessed in multiple studies. The problem is that most of these studies are themselves not terribly valid. One piece of evidence for the validity of SETs is that they correlate positively with grades, both within and between classes. Defenders of SETs argue that these correlations indicate that students learn more from good teachers. Critics argue that they indicate bias, because students are likely to evaluate teachers more positively if they expect to receive good grades. Since all students within a given class are exposed to the same teacher and the same teaching, the positive within class correlation is most likely due to bias. Between-class correlations – where the means of course grades are correlated with the means of teaching evaluations – provide more plausible evidence for teacher effectiveness. However, the assumption that students give more positive evaluations to teachers that grade leniently again offers a plausible alternative explanation.

> "The evidence consistently indicates that SETs do not measure teaching effectiveness or student learning"

To eliminate these ambiguities, educational psychologists developed the multi-section design as the gold standard for evaluating the validity of SETs. Assume that at a large university, introductory psychology is offered in many different sections, each being taught by a different teacher. Students are randomly assigned to these sections and they are all tested with the same centrally administered multiple-choice test at the end of the semester. Under these conditions, a positive correlation between the mean evaluation of the teaching in each of the different sections and the mean grade students received in each section would indeed be an indication of teaching effectiveness. Students should learn more in well-taught sections and also evaluate their teacher more positively than students in sections that are poorly taught. And the fact that they learnt more from these good teachers should also result in better grades.

Early meta-analyses, which statistically summarised the findings of such studies, have indeed reported moderately positive correlations between average SET scores in the various sections and section grade point averages (e.g. Cohen, 1981; Feldman, 1989). However, a recent meta-analyses by Uttl and colleagues (2017), looking again at these earlier meta-analyses, found that their results were mainly due to their reliance on studies that included few sections and their failure to control for such differences in sample size: they gave the same weight to studies with small as with big samples in their summary scores. Uttl and colleagues showed that if one weighted studies by sample size, the correlations between teaching evaluations and average grades were considerably reduced. However, the fact that numerous large sample studies had been conducted since these earlier analyses allowed them to conduct a much larger meta-analysis based on 97 multi-section studies. Uttl and colleagues found no significant correlation between the SET ratings and test performance. They concluded that their 'findings suggest that institutions focused on student learning and career success may want to abandon SET ratings as a measure of faculty teaching effectiveness' (p.22).

Although this meta-analysis is convincing, even more persuasive evidence for the invalidity of SETs comes from studies that use a new design, originally proposed by Johnson (2003) in his important book Grade inflation: A crisis in college education. Johnson suggested a brilliant and at the same time very simple solution – that the best indicator of what students learnt in an introductory course would be their performance in an advanced course that builds on the knowledge acquired earlier. In the meantime, six studies have been published that used this design; four conducted in the USA (Carrell & West, 2010, Johnson, 2003; Weinberg et al., 2009; Yunker & Yunker, 2003), one in Italy (Braga et al., 2014) and one in South Korea (Keng, 2017). Johnson found two positive predictors of future performance; namely, self-rated attendance and grading stringency. The greater the self-rated attendance and the tougher the grading in the previous course, the better students did in a subsequent course. In contrast, students' perception of instructors' knowledge and course organisation – the indicators administrators would pay particular attention to in evaluating teaching effectiveness – was negatively related to future performance. The other studies that only looked at average SET ratings found either no association or a negative association between average SET in previous classes and grades in subsequent courses.

From this I draw two conclusions, one with great certainty, and the other more tentatively. The evidence consistently indicates that SETs do not measure teaching effectiveness or student learning. The second and more controversial conclusion is that positive evaluations of teaching might sometimes even reflect poor rather than good teaching. As I will discuss next, there is some indication that teachers who require little work and give good grades, are likely to be rewarded with good evaluations. Whereas the first conclusion suggests that administrators should stop using these measures in their decisions about merit pay or promotion, the second conclusion suggests that they should have a closer look at instructors who

receive particularly positive evaluations. It is possible that some award-winning teaching results in little student learning. In the interest of full disclosure, I should mention that I never received a teaching award nor ever aspired to get one. I was always contended with the fact that my teaching was positively evaluated.

Wolfgang Stroebe *is Professor of Psychology at the University of Groningen*
wolfgang.stroebe@gmail.com

Poor measures of student learning

Every university teacher treasures students in their classes who are fascinated by their studies, and who work hard to learn their subject and to receive good grades. But what is it that students want? The publicly available website RateMyProfessors.com gives some indication.

Here, students can evaluate their university teachers and courses on dimensions such as helpfulness, clarity, and easiness. Helpfulness and clarity ratings are combined to form the score for quality. In one of the earliest studies of ratings of nearly 7,000 professors in Canada and the USA, Felton and colleagues (2008) found that quality ratings were fairly highly correlated (r = 0.62) with 'easiness' (defined as a course in which an A-grade can be achieved without much effort). This suggests that a large proportion of student raters prefer courses where they can get a top grade without too much time investment.

It has been suggested that these ratings are biased, because the website is likely to attract students from the extremes of the rating distribution. However, studies that compared the evaluations with ratings of the same professors on their university SET, found fairly good correspondence (e.g. Sonntag et al., 2009). This suggests that – if at all – SETs are certainly not a pure measure of teaching effectiveness. Instructors who do not impose workloads that interfere too much with their students' social life (or their ability to earn money to support their studies), and who also do not expose them to the risk of failing their course, can expect to receive more positive teaching evaluations than professors who require hard work and give tough grades. (I discuss additional corroborating evidence in my 2016 article.)

SET and grade inflation

The fact that many students prefer instructors who require little work and grade leniently has not remained a secret to university instructors. Although there have been few relevant surveys of university instructors even in the US, they indicate that most are aware of the association between workload and grading practices on teaching evaluation. Whereas some teachers might intentionally reduce course workload to improve their teaching ratings, such reductions could also be the unintended consequence of a decision to show more films and to spend more class time on explaining reading material. These processes might be at the root of a great paradox of higher education in the US, that grade point averages have increased for years, whereas the amount of time students devote to their studies has continuously decreased (Stroebe, 2016). This increase in GPA over an extended period without a corresponding increase in student achievements has been referred to as grade inflation (e.g. Rojstaczer & Healy, 2010). Grade inflation is particularly marked at private universities (e.g. the average grade for undergraduate courses at Harvard is now A -).

There is also grade inflation in UK higher education. Between 1994/95 and 2011/12 the proportion of first class degrees has doubled from 7 per cent to 15.8 per cent (Bachan, 2015), and has further increased to 26 per cent in 2018 according to *The Guardian*. Although it is less common in the UK for SETs to be used by administrators, one cannot but wonder whether the widespread use of SET by instructors does not contribute to this development.

A lost cause?

I've told of how numerous well-intended developments can ultimately result in unwanted consequences. SETs gave students a voice regarding the content of their courses. Since higher education is at least mostly about giving students a good education, it is reasonable that consumers of our educational offerings should have a say in evaluating the product. And after decades of evaluating university professors solely by their research output and publications, it was time to give teaching effectiveness equal status. In this context, adding the Teaching Excellence Framework to the Research Excellence Framework was a logical development.

So student evaluation of teaching may not be a lost cause, at least in principle. In practice, it is essential that we statistically control for all the known biasing factors, such as expected grading leniency and perceived easiness of workload. Even then, studies of validity are essential, and this might require a return to the drawing board. What are we trying to achieve with teaching, and what does excellent teaching therefore look like? Psychologists should be at the forefront of such debate.

Key sources

Braga, M., Paccagnella, M. & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review, 41*, 71–88.

Felton, J., Koper, P.T., Mitchell, J. & Stinson, M. (2008). Attractiveness, easiness and other issues. *Assessment & Evaluation in Higher Education, 33*, 45–61.

Rojstaczer, S. & Healy, C. (2010). Where A is ordinary. Teachers College Record, ID Number: 15928. www.tcrecord.org

Sonntag, M.E., Bassett, J.R. & Snyder, T. (2009). An empirical test of the validity of student evaluations of teaching made on RateMyProfessors.com. *Assessment & Evaluation in Higher Education, 34*, 499–504.

Stroebe, W. (2016). Why Good Teaching Evaluations May Reward Bad Teaching. *Perspectives on Psychological Science, 11*, 800-816.

Uttl, B., White, C.A. & Gonzales, D.W. (2017). Meta-analysis of faculty's teaching effectiveness. *Studies in Educational Evaluation, 54*, 22-42.

Full list available in online/app version.