

53 Powerful Ideas All Teachers Should Know About

Graham Gibbs

Idea Number 18, September 2014

You can measure and judge teaching

It is often argued that teaching is an art rather than a science, and that what works is infinitely varied, mysterious and not measurable. It is even more often argued that students cannot possibly judge teaching and that using feedback questionnaires involves merely a beauty contest, amenable to all kinds of skulduggery, making attempts at quantification of teaching utterly invalid and untrustworthy.

The issue of whether you can measure teaching, and exactly how to do it properly, has been the focus of more research than any other issue in higher education, and by a considerable margin. The leading researchers in this field, such as Herb Marsh, are the most frequently cited of all educational researchers on any educational topic. Much of this research was conducted decades ago and there is much less work on this issue today because the main conclusions have been broadly agreed for some time. This does not stop predictable criticisms of student feedback questionnaires being trotted out in public at frequent intervals, but they are invariably very poorly informed. So what is actually known about this issue?

Global judgements about whether teaching is simply 'good' are open to all kinds of varied interpretations by students about what 'good' means, and this leads to wide variation in

responses between students. If, in contrast, you ask students about specific teacher behaviours that are known to affect learning, such as whether their teacher gave them prompt feedback on their assignments, then students tend to agree closely with each other and such questions distinguish well between teachers, and can be trusted and interpreted.

There are many features of teaching which students can reliably distinguish and about which they make the same judgements about the same teacher on different occasions, and which they agree with each other about. Student judgements can be consistent and reliable. Teachers' colleagues can spot these features as well, and what is more they agree closely with the students' perceptions, making pretty much the same ratings and judgements as students do. Judgements made about these features are stable over time and discriminate well between teachers. The belief that students can only recognise the value of teaching in distant retrospect is contradicted by the available evidence: students do not usually change their minds later. If they think a teacher is rubbish they will usually still think they were rubbish 20 years later. Examples of students suddenly coming to the realisation, years later, that someone considered a rubbish teacher at the time was, in retrospect, actually quite good, may not be 'urban myths', but they are vanishingly unlikely.

53 Powerful Ideas All Teachers Should Know About

Graham Gibbs

Students can readily distinguish between teachers they like and teachers who they think are effective, and so this is not a beauty contest (provided you ask the right questions). There are some systematic biases in students' responses, but they are not disabling provided that measures are compared sensibly. For example comparing questionnaire scores for one teacher of a large enrolment compulsory course with no close contact with students with scores for another teacher of a small optional course with generous teaching resources, would not be fair. Using questionnaire scores in personnel decisions is risky unless accompanied by a range of other evidence and interpreted with common sense.

Well developed questionnaires, such as the SEEQ (Student Evaluation of Educational Quality) produce scores in relation to those aspects of teaching that students can reliably judge (such as about feedback). Most of these scores have been found to link reasonably closely to all kinds of outcomes of good teaching, such as student effort, student marks – so measures can be valid as well as reliable. It is often argued that students judge to be good those teachers who award them high marks. This problem is overcome in studies of large enrolment courses where many teachers each teach their own small group in parallel, with all students then sitting the same exam marked independently. It is usual to find differences in average student marks between groups that can be attributed to measurable differences in student perceptions of the teachers of these groups – and their teachers did not mark their work, and students rated

their teachers before they were marked. The most widely accepted interpretation of these 'multi-section' studies, as they are called, is that the best teachers, as judged by students, produce the best student performance.

It takes many years of rigorous research to develop a measure such as the SEEQ. Most student feedback questionnaires used in the UK are 'home made' and lack any evidence of reliability or validity, include variables known not to be linked to student performance, and do not distinguish well or consistently between teachers or courses. Many are likely to be both untrustworthy and uninterpretable and deserve much of the criticism they receive.

Can you measure teaching across a whole degree programme?

What is driving the current HE market is measurements about students' experience of degree programmes, rather than of individual teachers – and this involves a great deal more than aggregating measurements of the quality of individual teachers. Research has identified the characteristics of degree programmes that are associated with the greatest learning gains. The overall effect of these influential variables is to improve student engagement: both its quantity and its quality. As engagement predicts learning gains, degree programmes with the best teaching can be argued to be those that achieve the highest levels of student engagement. An American questionnaire, the National Survey of Student Engagement (NSSE) measures engagement and predicts learning gains well. If you use the NSSE to

53 Powerful Ideas All Teachers Should Know About

Graham Gibbs

identify teaching quality problems, and use the research evidence to select and adopt alternative approaches to teaching that are associated with better engagement, then student engagement improves and learning gains improve. The NSSE works so well as a measure of teaching, and to improve teaching, that 800 US institutions voluntarily pay to use it and it is used widely in other countries such as Australia. The Higher Education Academy has piloted a short version of the NSSE in the UK. In contrast the score from the use of the UK's National Student Survey (NSS) that is most commonly cited is that concerned with 'satisfaction', which does not predict learning gains and cannot be considered to be a valid indicator of teaching quality. All kinds of things result in at least some students being highly satisfied (such as the course enabling them to get adequate marks without having to work at all hard) which have nothing to do with good teaching or which might even indicate bad teaching. There are several questions on the NSS which are likely to be valid, such as those concerning feedback, but we currently have no evidence that any of the NSS is valid, in the sense that it predicts how much students will learn. This does not mean that measuring teaching is not possible, simply that the UK uses the wrong measures.

Can you judge teachers when making promotion decisions?

When academics compete for promotion on the basis of research achievements this does not involve measurement, but judgement. And judging research is not an exact science, as

evidenced by reviewers of my articles who have had diametrically opposed opinions. But at least, when comparing researchers for promotion, expert judgements have already been made by reviewers about articles and grants, and the overall judgement of the individual by a promotions panel is based largely on these prior expert judgements. The main problem with judging teachers for promotion is that expert peer judgements of their teaching have not normally been made at an earlier stage. In research universities where judgements of teaching are central to promotion decisions, such as at Utrecht or Sydney, teachers have to collect elaborate evidence over time, just like developing a research cv. It has taken departments in these institutions a decade or more to work out how to judge such teaching evidence consistently and with confidence. Both standards and expectations have risen as they have got better at it. Just as inexperienced journal reviewers sometimes get it wrong, so inexperienced judges of teaching sometimes get it wrong. Judging teaching is difficult and time consuming, but probably no more difficult than judging research, and you can learn to do it better.

So it is possible to measure teaching both for individual teachers and for programmes, and it is possible to judge teaching for promotion decisions. Much of the usual criticism is justified, however, for many current practices.

This '53' item has been developed from an article by Graham Gibbs published by Times



53 Powerful Ideas All Teachers Should Know About

Graham Gibbs

Higher Education, and their permission to use text from that article is gratefully acknowledged.

To comment or contribute your ideas, see SEDA's blog: thesedablog.wordpress.com