53 Powerful Ideas All Teachers Should Know About

Graham Gibbs

SEDA
STAFF AND EDUCATIONAL DEVELOPMENT ASSOCIATION

Idea Number 43, February 2016

Title Most assessment involves (unreliable) professional judgement - and is all the better for it.

Over 40 years ago Roy Cox summarised a wealth of research evidence about the reliability of marking in universities. What the evidence showed was that much marking, across a wide range of disciplines and types of assignment and examination questions, was extraordinarily unreliable. Markers disagreed with each other to a startling extent. In some studies, who the marker was contributed more to variance in marks than who the student was. In my own studies of marking, involving multiple markers, most individual student final year project reports received marks that varied by one or two degree classifications. Agreement between markers was rare.

I have sometimes experienced the same phenomenon when submitting articles to journals – receiving one reviewer's wholehearted praise and support and another's implacable opposition, but on entirely different grounds. When the article has been submitted to a third reviewer this has simply introduced an additional and unique perspective and left the editor with the problem of how to proceed by making a fourth, subjective, judgement. It isn't always like this but it has happened to me often enough to make it clear that the statements of criteria listed in journals' instructions to authors cover a multitude of sins.

Thirty years after Cox's summary the then President of the British Psychological Society conducted a series of marking reliability studies in his own degree programme, to see if this unreliability was still the case despite modern attempts to specify educational goals and assessment criteria, and even though Psychologists are supposed to understand psychometrics, testing reliability, testing bias and so on. Surely, with quality controlled specification of assignments and criteria, psychologists could act rationally and eliminate unreliability from their marking! Not so, it was found. What is more adding second markers made things worse and adding external markers made things much worse – they simply brought new and different standards with them and added to the randomness of the whole process. You could average out multiple markers' scores, which tended to produce a narrow range of middling marks, but the central problem was that markers differed in what they were doing when they marked.

I once undertook a study of standards in marking in a context where student numbers had soared, to see if, over a decade, standards had gone down. A collection of final year projects that had been archived by the department from a decade before, and from the year of the study, were marked by a series of internal and external markers, and the

marks compared. At first it looked as though everything was fine. Contemporary students' average marks were very similar average marks of the students of ten years before, and the students' reports from ten years before were given pretty much the same marks, on average, by contemporary markers as they had been given by the markers of ten years before. In terms of average marks standards appeared to have been maintained. However on closer inspection equivalent average marks were only achieved by the marking being effectively random. Much the same average and spread of marks was achieved, but with the rankings of students differing wildly between markers. With such extraordinarily unreliable marking it was impossible to tell whether standards had changed. I went further and asked the markers to go back and rate the reports on a series of criteria. It turned out some markers' marks were closely correlated with their ratings on criteria concerned with the academic subject matter (e.g. "Made good use of contemporary theory") while other markers' marks correlated with a completely different sub-set of criteria that concerned generic issues (such as writing and referencing). What is more if a marker's marks were determined more by generic criteria, then their marks tended to be higher. Not only were marks effectively random, but hidden from view the markers differed widely in terms of what they were looking for and valuing, what their marks were based on, and what standards they were adopting. Needless to say the formal External Examiners were not able to spot any of these phenomena and had given the Department's a clean bill of health for both iterations of the examination system a decade apart.

You might protest and claim that in your own degree programme second markers do not show such wide variation from first markers as the research evidence suggests. This is indeed often the case. The studies cited above were in 'controlled' situations – markers did not know what others were doing and did not talk about it – all marking was 'blind'. Most second marking, in practice, is surrounded by discussion of examples and differences of view, and takes place in contexts where markers have been marking and talking to each other about their judgements about familiar forms of assignment, and even about familiar questions, for some years. Conversely if new markers are brought in they tend to 'get it wrong' and it may take a couple of years' practice before they gradually align what they are doing with their colleagues. Doctoral students' marking is notoriously out of line with expectations. A psychology degree programme I knew at an institution that had expanded student numbers recklessly quickly had to draft in part time markers to cope with the assessment load and their marks were so all over the place that they had to get the over-worked full timers to do all the marking again in order to pick up the pieces.

It is the, often informal, 'community of practice' work that goes on that helps to reduce unwanted unreliability. In the Law Faculty at Oxford they exploited this phenomenon in order to reduce the amount

of double or even triple marking that traditionally took place when marking finals papers. Markers would sit in the same room and compare notes about student answers they were struggling to grade. Issues that were specific to a question, set in the context of the Faculties' long standing values and ways of doing things, were talked about as the marking progressed, gradually bringing different markers' perceptions and judgements closer together. In this way unwanted variations between markers were reduced and they ended up only double marking exam answers that were very close to degree classification boundaries, where the final mark mattered more. The effort put into aligning their judgements cost less time than double marking everything.

Interestingly at a number of institutions there is no student appeal allowed against marks, only against failures of procedure. Academics' marks are (in private at least) considered so subjective that there is no point in discussing their variability and no way to resolve differences of view by bringing in additional markers. Academic judgements are final - or everyone would go mad.

It might also be argued that such wild unpredictability of marking only happens in Sociology or other text-based subjects, and with open-ended essay-based student answers. I came across an amusing staff development exercise concerning assessment that was used to disabuse university teachers of this view. In it all those in the workshop were asked to mark a long division sum and allocate a mark out of 20. The answer to the sum was correct but there was a mistake in the working out. Over the years the person who invented this exercise had collected all the marks hundreds of academics had given this long division sum – and every mark between 1 and 20 had been allocated at some point (though curiously never zero – perhaps all academics are soft at heart). I recall 14 being the most common mark, then 19, then 7, and so on. It was extraordinary. When questioned, some markers said the answer was right and that was all that counted, but they never gave any student full marks for anything and so had given 17. Others said that the student must have cheated to get the right answer (as the workings were faulty) and so gave it 1. Others said that it wasn't a very difficult sum, and so could not in all fairness give it more than 11... and so on. The most extraordinary range of rationales were trotted out to justify absolutely any mark. I am sure that subjects that use quantitative questions with correct answers have, over the years, and through encountering anomalies and talking about them, resolved many of these differences of perspective and values - but almost certainly not all of them.

However the opposite of relying on professional judgement is, in my view, even worse. Attempts to nail everything down with criteria and marking schemes, and to devise 'objective' testing, produce at least as many anomalies and also tend to produce crushing uniformity. Some readers will have come across the example of the school science question that asked how to measure the

height of a tower with a barometer. The 'correct' answer, and the only one that deserved any marks at all according to the marking scheme, was to measure the barometric pressure at the top and bottom and use a formula to calculate the height difference from the pressure difference. But students answering this question came up with much more interesting answers, including dropping the barometer from the top and timing how long it took to smash into the ground, lowering the barometer on a piece of string until it almost touched the ground and swinging it like a pendulum and timing how long each swing took, or even just measuring the string (in sophisticated versions taking Hookes Law into account). If I were interested in developing flexible scientific thinking then I would consider these would be very good answers, or at least perfectly adequate, but the students who thought them up all scored zero.

A crucial issue in the 'barometer' story is who set the question and who specified the 'right' answer and its associated marking scheme, and why. And if different teachers had set it or specified the marking scheme would they have done it differently, for different reasons, and so produced different marks? The reality, of course, is that there is no such thing as 'objective' assessment. It is always a matter of professional judgement – at the design stage if not at the marking stage.

I'd be happy to rely on professional judgement in marking – but I'd hope the markers talked to each other about how they derived their marks, and kept talking, for years.

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