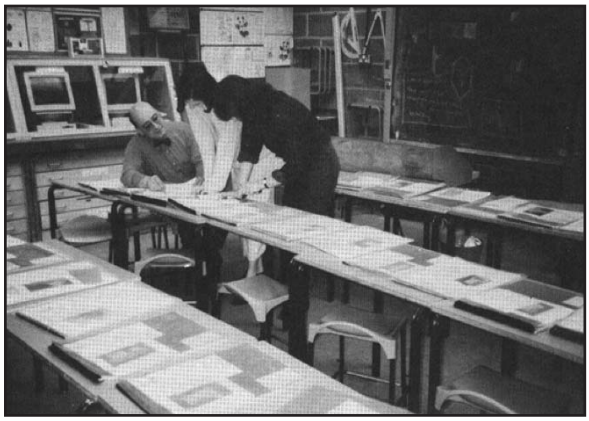
Holism and the Challenge of Teachers' Judgement

Prof Richard Kimbell, Goldsmiths, University of London

In 2003 I wrote a DATA editorial entitled "The whole is greater (and more important) than the parts". I was referring there to a holistic sense of the D&T curriculum, but the same is true – I believe – when thinking about approaches to assessment.

Typically, regardless of the purpose of any assessment, what we do when making judgements about learners' work is to identify a set of qualities that are important to us and then seek to gauge the fit between these qualities and the work itself. In summative (e.g. GCSE) assessments, these qualities are explicitly identified as assessment criteria and have numbers attached to them so that we can 'score' learners' performance against each of them. We then add up all the part scores and arrive at a total score that results in a grade being awarded. The question that we must answer however, is whether this added-up collection of parts is the same thing as 'capability'. I am on record as doubting this; see for example chapter five of 'Assessing technology' (Kimbell 1997), in which I have a go at atomised assessment generally and – specifically – the lunatic regime that was thrust upon us in the early years of the NC.

The problem however is that whilst there is very good reason for us to draw up a list of qualities that are important in design and technology, such lists tend towards atomisation rather than holism. I know that it's helpful to think about the various ways in which learners might display strengths and weaknesses – and in which they might be encouraged to grow and develop. So what are we to do with such criteria lists if we are not to 'score' them... thereby accentuating the atomisation problem?



As a start, think about what we typically do when asked to assess a batch of portfolios. I apologise for the quality if this old 1980s photograph – but it illustrates the point rather well. The teachers started their assessment process by laying out the portfolios across three lines of tables – with the 'best' at one corner of the room and the 'weakest' at the opposite corner. They then examine each in turn and go through the detailed process of ascribing scores for each of the criteria.

My question is…. "how do they know which are the best/worst before ascribing the numbers"?

The answer of course is that having taught them for ages we hold a strong intuitive sense of each learners' capability and therefore just 'know' which are the most and least capable. So ascribing numbers becomes a process through which we attempt to validate our primary judgement with some crude arithmetic. And I suspect that if the arithmetic does not come out as we would expect, we are more likely change the numbers than to change the original rank order that we started with. We typically trust our judgement more than our arithmetic. And rightly so.

So there… in case you missed it... I have used the dreaded words 'rank order'. For when we lay out the portfolios using that initial judgement, we are ranking learners in relation to each other. But surely this is just horrid normative assessment? Isn't GCSE supposed to be about objective, criterion-based judgement? The reality is that ALL assessment requires both norms and criteria. Judgement always – inevitably – requires us to compare one thing with another thing. That is why exam boards provide exemplars that illustrate (give us a comparison point for) what is meant by grade A or C of F. The reality however about exemplars is that nothing is directly comparable to the piece I am trying to mark because of the countless ways of being good/poor at D&T (or anything else for that matter). So we are forced to use imagined and inferred comparisons that are fragile and that result in the reliability of the assessment process being barely adequate.

So how might holistic assessment help with this problem? Surely if atomised, criterion-based assessment is unreliable, then holistic judgement – which must involve the interaction of lots of these criteria – is going to be even more flaky. Well... interestingly... not so.

The literature has for ages attested to the fact that teachers' rank-ordering of their class is remarkably accurate. Its one of the forms of assessment that is astonishingly reliable. And for two reasons. First, we know the learners well and second, the comparison that we have to make are very direct ones. I don't have to compare John's portfolio with an abstracted set of criteria/exemplars that might or might not fit with his work.

Holism and the Challenge of Teachers' Judgement

Rather I have to compare John's portfolios with Henry's. They are both there in front of me and my resulting judgement is far more likely to be accurate.

When the teachers in the photo made their initial judgements they were – of course – making such a holistic judgement. They would probably have been able to articulate a few criteria that are relevant – but they were standing back from those details and were able to say that 'overall' this piece is better than that one. Its a very natural thing to do. The inaccuracy arises when we go on to assert that this portfolio is 'worth' 122 marks out of 160 (or whatever).

It is this process of teachers making holistic judgements that we have been studying – and developing – in the e-scape project (see Paul Clewes 'The Great Escape' in the latest DATA practice).

Essentially the e-scape system allows learners to create real-time e-portfolios of their response to a task. These e-portfolios are visible at a glance... the software jargon for it is 'blink-ware'. Each of the thumbnails is a live link and a single mouse click on it brings the image (or text) full screen. The portfolio is based on a rich media mix of text, drawing, photo, voice files, and video.

And this makes it possible to employ a means of assessment that uses direct comparisons. Teachers are presented with two portfolios (A and B) and have merely to decide which portfolio reflects the more capable performance. This sounds like classic norm referencing, but in reality our teams of teachers are trained to look for particular criteria in the work. We need to do that so we can describe what you mean by capability. But having discussed the qualities, we don't want assessors to 'mark' them, but rather to make a balancing judgement about whether portfolio A or portfolio B contains a better mix of qualities that amount to capability. So the exercise quite explicitly requires a mixture of norms and criteria in arriving at a holistic judgement.

In the research just completed (with 350 portfolios), 28 teachers each made 120 paired judgements, and – as if by magic – the data emerges as a complete rank order which can then be analysed into grades of performance. From the point of view of our teacher/assessors, it really is assessment without numbers. For those interested in how this works, you can read the details in the e-scape reports in the Goldsmiths/TERU website.

But at the heart of it lies the challenge of whether teachers can grapple with the idea of 'capability' as the amalgam of a group of criteria. And whether they can assess portfolios reliably using this holistic notion. The fact is that they could and did – with a reliability coefficient of 0.95 – which is far higher than normal GCSE reliability statistics. When the teachers were asked to reflect on what is an utterly different model of assessment to the one they have been brought up on, they made all kinds of interesting observations.

GCSE marking relies heavily on a tick box assessment of a pupil's work. It can be frustrating when confronted with an excellent piece of designing and making that does not meet the exam board's criteria. Too often the linear pattern of coursework requires the assessor to jump back and forth to find the marks that a student deserves. The e-scape judging is so simple in comparison.

It gives more appropriate results than atomised approaches which can lead to inaccurate overall assessment especially when the overall attainment is more than the sum of the parts. This often happens when the various elements of a designing process come together in a successful outcome that outstrips the quality of work in any (or all) the parts of the process.

Given our traditions and practices in D&T, I was not surprised that our teachers had no difficulty with the idea of capability-based assessment, and indeed that they were enthusiastic about it. But e-scape has been venturing into other territory – into science and geography and English. What would these teachers make of it? The holistic use of a descriptor is very different, but also very powerful. It forces all to really reflect upon the geography, rather than tick boxes in a mark scheme.

I would like to believe that all teachers have this intuitive sense of what amounts to excellence in their discipline. Just as the teachers in the 1980s photo displayed. But the difference now is that we are beginning to construct a methodology that respects this holistic teacher judgement and enshrines it within a rigorous and authoritative framework for awarding. Watch this space.

Kimbell R (1997) Assessing technology: *International trends in curriculum and assessment*. Open University Press, Buckingham.

r.kimbell@gold.au.uk