The Transient and the Timeless: Surviving a lifetime of policy and practice in assessment

Prof Richard Kimbell, Goldsmiths, University of London

RESEARCH

A personal retrospective

1960s

In1959 I half-passed my 11+ and went to a technical school in Kent. In 1964 I took O levels and in 1966, A levels in metalwork and technical drawing. These examinations were unimaginatively constructed – amounting to tests of theoretical knowledge (theory paper) and practical skills (the practical). I was good at them.

During and after the war, my teachers had been responsible for setting up production lines in the school – manufacturing parts for the ships being maintained in Chatham dockyard. They knew about precision, we had lots of practice and we ended up pretty capable draughtsmen/craftsmen (it was of course a boys school).

I went to Goldsmiths to train as a handicraft teacher, and I was seriously shaken up to discover designing. I was used to being given engineering drawings of tool clamps and drilling jigs that I just had to make. Designing things for myself was a revelation. I did not at first enjoy it. I wanted just to revel in my skills that were better than the norm in the first year group because of the experience and facilities to which I had been used at school.

At the end of my first year I had one of several ‘road to Damascus’ experiences at Goldsmiths. It was in the end of year examination. We had to design a folding chair (for an art gallery) in a week (with one tutorial) and then make it unaided in two days (12 hours). Mine was in square tubular steel and folded really flat. It was a bit heavy but it did work. What shook me up was the astonishing variety of designs and I remember sitting round at the end, knackered, looking at all that stuff thinking “all this is completely new...it just came out of people’s heads and ended up as real”. I was electrified by the idea. At the end of my final year (July 1969) I was sitting with a bunch of mates at an all-nighter of TV, watching the moon landing. I thought it again then, all that astonishing stuff just came out of people’s heads.

What stayed with me from that experience was the power of ideas. A notional design process, articulated in a sketch book and crystalised on a drawing board, and the practical capability to make them real.

1970s

After an extra year at Goldsmiths doing the BAEd I started teaching in Devon. I was on my own, creating a new department and it was fun. I could do exactly what I liked. The county adviser dropped in now and then but generally left me to my own thing. We made swimming pool seating using cast concrete units from chipboard moulds…turned boomerangs into offensive weapons …made stained glass windows for the church from cast resin with dyed wood-shavings embedded in it…and survived on a beach for a week in a Neolithic camp; making stuff from what we could find.

In 1972 I started a GCE group on George Hicks’s London O level D&T. The revelation here was the pre-practical test. There was theory and a practical but George managed to get London to embrace a design paper without calling it design. In the pre-practical paper, students had several days to design and make a component of a product that was to be fitted to something that became the three hour practical test. I recall a thoedolite stand (the three hour practical) and students had to design a levelling system to fit to it. At the same time Eggleston, at Keele University, was shaking the trees with his CSE ‘Course of studies in design’ and his assessment system was astonishing. Entirely process driven and I think you will find it familiar. In 1973 I started a CSE group on this and was immediately thrust into juggling with the GCE/CSE nightmare.

In 1979 Margaret Thatcher was elected.

1980s

If the 1970s had been a decade of expansion and proliferation, the 1980s was one of constriction and retrenchment. Thatcher created the Secondary Examinations Council (SEC) requiring all subjects to define themselves in ‘National Criteria’. Only one set per subject. So only one thing could be taught under the name of design and technology. All the existing courses in crafts, design, technologies, technical drawing and graphics were to be squeezed and massaged into a single pot. This exercise in definition led to some of the most ferocious debates I have ever experienced. There was blood on the walls in Notting Hill Gate. Having fought our way through the National Criteria exercise, we then embarked on the enormously complex process of establishing the GCSE examination itself (GCE+CSE= GCSE) with all the attendant arguments about course content, examination procedures and educational standards. GCE was about academic rigour and university entrance. CSE was much more tuned in to learning styles and authentic learning experience, and it embraced coursework. By 1988 the first cohort of pupils was just completing its GCSE courses and taking the first round of examinations set by the newly merged Examination Groups. It was clearly time for another upheaval, only this time the upheaval was absolute. We embarked on the national curriculum. Hargreaves (1989) observation that 'it would be no exaggeration to say that the 1980s has been an era of assessment-led educational reform' was the understatement of the decade.

Throughout this time I was continuing as an Oxford A level Design examiner. The combination of autonomy for teachers to design their own courses, linked to a sensibly flexible assessment regime, managed to create an oasis of individuality and innovation in a desert of standardisation and control. And all this control originated from Thatcher the de-centraliser/free marketer. She nationalised the curriculum whilst de-nationalising everything else. Funny old world.

1990s

The NC took its shape through the Task Group on Assessment and Testing (TGAT) and through the Statutory Order for Technology (DES 1990). Within weeks of the publication of the Order teachers were struggling to assess pupil performance in technology against 149 'can-do' statements of attainment set at ten levels. It was a process that ran completely counter to the whole of teachers' experience of assessment in schools and it was completely barmy. On top of the teachers school-based assessments came the official testing in the form of Standard Assessment Tasks. Initially for pupils at age 14, then also for age seven, these new assessment instruments made yet further demands on teachers, testing their understanding of the new definition of technology. For many primary schools this was their first encounter with technology.

Within months of trying to teach and assess technology as defined in the 1990 version of the national curriculum teachers were told they were getting it all wrong. Her Majesty's Inspector (HMI) wrote critical reports and all the special interest groups that felt they had been ignored or marginalised by the 1990 formulation of technology were out sharpening their knives and grinding their axes. It was a period of generally ill-informed and very public blood-letting. In the end it was the teachers who put a stop to the assessment madness by their astonishingly solid refusal to have anything to do with the 1992 and 1993 SATs. But by that time, the problems of technology were being blamed *not* on the assessment system but on the definition of the subject in the Statutory Order. It was time for yet another re-write. Over the following two years another three versions of technology were written and debated before we finally ended up with a new Statutory Order (DES 1995). And then another new one in 2000.

2000

At the Design and Technology Association's Millennium Conference (2000) I presented a paper entitled 'Creativity in Crisis' and it surely was after ten years of curriculum strangulation allied to the pernicious influence of Ofsted. Schools were running scared and appeared to be prepared to do anything (however daft) in pursuit of their treasured A-C GCSE pass rate percentages. But the sheer ferocity of the compression and standardisation process, that had started in the 1980s and that had then been brought to fruition in the national curriculum of the early 1990s, was bound to result in a backlash. Pendulums have a remorseless habit of swinging, and as we launched into the naughties, more and more people were prepared to talk about the constricting effects of curriculum. Talk of creativity re-emerged, and we saw initiative after initiative aimed at encouraging it.

The Key Stage 3 initiative was one such, that was specifically encouraging teachers to try something new. But in the assessment world we are still in danger of moving backwards. The brave move towards 14-19 Diplomas was an important curriculum initiative that might yet be still-born, and the election has not made things any clearer. But whilst brave and important as curriculum, its assessment leaves so much to be desired. All the indications so far are that the associated assessment will follows the ghastly model of 'outcomes' atomising the learning process and being scored on ever-expanding tick-sheets. In reality most of the diploma assessment can be internal to the schools and one would hope that some schools will be brave enough to challenge this national curriculum-ish model. But the signs are not good. It needs teachers with real confidence to stand out from the herd.

Assessment in transition

How can it be that school curriculum policy and school-based assessment policy, can be subject to so much conflict and change? Why can't we just work out the best solution and then leave schools and exam boards to get on with it? Why does education policy (and hence practice) have to suffer the dramatic upheavals that have been so evident throughout my life in schools?

It might help towards an answer to that if we could identify some of the major variables that underpin these changes and that lie littered across time. Any one of them can be the source of some serious disagreements about what counts as good assessment.

Conflicting philosophies

Philosophers have always speculated about right and wrong; the good and the bad; the just and the unjust. By what means can we tell them apart? Bentham and his Utilitarians took one view, and it was very different from that espoused by Plato in his world of forms or Rousseau and his concept of natural growth. Who is to say where the *real-truth* lies?

General philosophies are readily translated into models of assessment. The empiricists, like Hume and Skinner were comfortable with assessing components of knowledge and building from these components towards the concept of 'mastery'. In performance assessment terms, theirs is the world of psychometrics. The rationalists, like Descartes or Piaget were more interested in looking at extended performance and were prepared to credit many varieties of excellence. Critically however they wanted to assess the extent to which learners could make use of their learning experience when tackling new problems. And if they can't, how can we describe it as 'learning'? And then we have the socio-culturalist like Marx, Vygotsky and Luria who are more interested in assessing the learners' participation in the social processes and practices of learning. As they create their portfolios are they sufficiently self aware that they can take their own part in the assessment process?

These are just a few of many philosophical positions that may be adopted in the realm of assessment. They all make sense in terms of their own inner logic but they do not comfortably acknowledge each other.

The scale of the enterprise

A further variable in the evolution of assessment concerns simply the question of scale. At a recent conference in Australia I heard a delegate from China describing their early planning for assessment and they were modelling ideas on the assumption of 10s and 100s of millions of cases.

But we don't need to go that far to see the effects of scale on the evolution of assessment practices. The original concept of the lone scholar, producing a thesis and being examined in a viva-voce (the living voice) examination typifies a very ancient tradition that is retained to this day for very high status assessment. It's a very personal kind of examination tuned exactly to the individual priorities of the work of the student. Its all about trying to understand what has been done. And, even though what has been done in this case is entirely different from anything that others have ever done before, it does not prevent us from saying that YES this student has achieved at a sufficient level to pass.

The Victorians in Britain were the first to create a mass system of education specifically to feed the demands of empire. So many were needed to be able to read, write and reckon, that it became necessary to school everyone. And that resulted in the first attempts at mass-assessment with lines of desks and standard tests. The year groups in schools were even called 'standard 4', 'standard 5'. In America this same issue provoked a different strand of assessment thinking in 1914 when the army needed recruits for the first world war. With typical American ingenuity they saw the need to test millions of men as an opportunity to standardise and automate the process. And so was born 'multiple-choice' testing and machine marking, that remains to this day the basic mode of operation in schools throughout the USA and its spheres of influence (e.g. Taiwan).

The science of assessment

As the scale of the enterprise builds and therefore becomes more significant in the life of the nation, it is important that the products of assessment should be correct. At least they should be seen as correct in terms of whatever prevailing philosophy is the order of the day. There are at least two good reasons for this demand for correctness, first concerning simple fairness in the treatment of all students, and second in terms of efficiency and cost effectiveness. If it costs a lot of money to conduct assessments we don't want rubbish data at the end of the process. And this inevitably creates the need for the evolution of a science of assessment.

In terms of school-based assessment, this science has now resolved itself into a form in which three factors typically play off against each other.

Assessments should be *valid* – meaning that the test/assessment activity should be an appropriate test of that quality in question and not of some other one. As an example, language (French) *conversation* cannot validly be tested in a written mode.

Assessment should be *reliable* – meaning that whatever judgement teacher X makes about the work must also be replicated by the judgement of teacher Y. So reliability might be thought of as 'repeatability'.

Assessment should be *manageable* – meaning that whatever the mode of examination it must be do-able with the number of candidates and the normal facilities available. This manageability criterion was very significant in the early discussions about the availability of calculators in maths exams.

Typically these three arms of the science play off against each other. VERY reliable assessments (e.g. machine-marked multiple-choice tests) can often be challenged for their lack of validity, whilst highly valid assessments (e.g. coursework portfolios with a viva-voce examination) might be challenged as being both unreliable and unmanageable.

It is interesting to note that when I was an Oxford A level examiner in the 1970s and 1980s we conducted a viva with every student. We had a small team of regional examiners who travelled to every school and spent time with every student – and the whole process was seen by the schools and the examiners as an important validation of the quality of the work. But inevitably as the candidate numbers grew, it became prohibitively expensive and therefore unmanageable.

Knowledge – skills – processes and capability

A further source of confusion lies in the substance of the 'stuff' to be assessed. In the 1960s I was typically tested on the knowledge I could recall (in theory exams) or the skills I could exhibit (in practical exams). The transition to design assessment raised the whole debate about design *process*, and it was much to our credit in design and technology that we have grappled so effectively with the challenge of process-centred assessment.

What we have been less good at (in my opinion) is in dealing with the consequences of the argument about process. The heart of any set of process skills lies in their *performance*. I might have good research skills (finding out), and good development skills (growing an idea), and good modelling skills (lashing up prototypes), and good reflective skills (reviewing performance). So the idea of 'performance' could be seen as comprising a number of sub-elements of performance. To avoid confusion we use the term *capability* to describe the overall performance. But where should we place the locus of assessment? Do we score all the bits and 'add-up' the performance? Or do we score the performance and then disaggregate the elements of it?

Holism and the importance statements

In England, as part of the 2000 National Curriculum the QCA published for the first time an 'Importance Statement' for each subject. This was a whole statement about the importance of the subject in the curriculum…and the Geography one is quoted here:

> The importance of geography
>
> The study of geography stimulates an interest in and a sense of wonder about places. It helps young people make sense of a complex and dynamically changing world. It explains where places are, how places and landscapes are formed, how people and their environment interact, and how a diverse range of economies, societies and environments are interconnected. It builds on pupils' own experiences to investigate places at all scales, from the personal to the global.
>
> Geographical enquiry encourages questioning, investigation and critical thinking about issues affecting the world and people's lives, now and in the future. Fieldwork is an essential element of this. Pupils learn to think spatially and use maps, visual images and new technologies, including geographical information systems (GIS), to obtain, present and analyse information.
>
> Geography inspires pupils to become global citizens by exploring their own place in the world, their values and their responsibilities to other people, to the environment and to the sustainability of the planet.

In just 160 words it aims to capture the essence of why students should study geography, and it gives anyone who reads it a good idea of what would distinguish a 'good' geography student from a 'poor' one. The Importance Statement has many uses. Once it has been accepted by all the relevant stakeholders, it provides a warrant for designing the syllabus, for choosing what must be included and as a basis for making the difficult decisions when there is too little time to include everything that everyone would like to see in it. Parents and prospective pupils may find it more useful that the detailed syllabus when deciding which subjects a child should study, and professionals outside the school system can see what knowledge and skills the course is meant to develop in the pupils. For teachers it provides a general criterion that they can use to judge the value of any activities they might consider using with a class, a constant reminder of what they are meant to achieve with their pupils, and a simple statement of what the examinations or tests will deem important.

> Assessors, it follows, must also use the Importance Statement in their work. They must ensure that their tests do indeed seek evidence of what is *agreed to be important*, and that they do indeed give credit for

> evidence that students' have learned what the Importance Statement defines as important. For this assessment function, the brevity of the statement is important; there are other documents that assessors might use to guide them in creating tasks and marking responses – aims, objectives, learning outcomes, grade descriptors – but these are often too detailed, and too list-like, to keep the assessor's mind focused on what the overall purpose of the test is – measuring what is important. The Importance Statement expresses the general principles for judging importance, and is brief enough for every participant to keep it clearly in mind, but also has enough detail to be useful in practice. A good Importance Statement will ensure that validity is maximised at the start of the assessment procedure.
>
> (Pollitt 2009 p4)

The recurrent underlying driver here is apparent. It is towards holism. The Importance Statement provides a succinct overview rationale; learners' capability provides us with a measure of their performance in this domain and teachers are adept at making these judgements holistically. Indeed, the behaviour of teachers coping with NC statements of attainment provides a ready indicator of the primacy that attaches to holism.

I watched two teachers in 1991 grappling with the assessment of a groups' work in the early days of the NC assessment arrangements. They scored all the work against all the SoA, which was an astonishingly atomistic and therefore long-winded process, added them all up and arrived at a rank order for the group. But then they stood back making comments like "Jane can't be better than Peter and both are better than Paul". The teachers knew exactly who was relatively strong and who was relatively weak and they were not convinced by the aggregated scores. So they went back and changed the SoA scores and re-calculated so as to arrive at the 'right' answer.

There is no doubt that teachers are far more effective as assessors when they start with holistic judgements and only thereafter disaggregate the elements of performance for diagnostic and developmental purposes. The Oxford A level Design assessment schema provides an interesting model of how this worked in the 70s and 80s and interestingly it is a model that is now reflected in our undergraduate and postgraduate student assessments at Goldsmiths.

Only the brave…

The 1970s was a decade of innovation. All kinds of odd stuff was going on. Some of it has rightly disappeared in the interim and some has remained and completely changed the way we see ourselves and our work. In the 1980s the freedom for teachers to innovate was ruthlessly squashed and in the 90s a rigid orthodoxy was imposed that has only gradually eased.

It has always been the case that some teachers and schools will be more risk-averse than others. When the crap is flying most will keep their heads down and play the game, putting their trust in the age-old dictum 'he who survives lives to fight another day'. Other teachers and schools will be more challenging, and whilst some will come a cropper, some are likely to be the source of new ideas that catch on and that completely re-formulate what might be possible and desirable in schools.

One thing is clear. All the literature on creativity asserts the importance of *self-confidence* in those who would be innovators. Jeffrey & Woods explored children's attitudes towards creative work in the classroom. The study

> ...draws attention to the need for trust in a creative classroom. The emotional climate of the classroom needs to offer each child personal confidence and security
>
> (Jeffrey & Woods 1997 p15)

The need for this trust relationship derives, of course, from the fact that creative acts are risky acts, and no-one will go out on a limb and take chances if they believe that should they fail, they will suffer serious penalties.

> a powerful theme in our own research was the belief that self-esteem and self-confidence must be nourished in order to be creative
>
> (Craft A 1997 p 83)

In his slightly weird Dome millennium event Tony Blair enthused about the need for a creative Britain and he too recognised the importance of confidence.

> Our aim is that risk-takers are rewarded. Let us believe in ourselves again. Britain's future depends on those with confidence, who take risks, like the creative talents we celebrate here today. They are the people that Britain needs in the next century...those who have ambition for our country.
>
> (Blair 1999)

I have never seen any statistics about the relative size of each group; the rule followers as against the innovators. Some people I suppose will always be rule-followers and some will always be rebels. But many, I suspect, are influenced by the mood of the times. The 70s encouraged everyone to be an innovator. In the 90s only the hardiest of rebels put the heads above the parapet.

Whatever the numbers, it remains true that we are taken forward in our thinking and our practice by that group of individuals who strive to do it differently. We need them.

Assessment as judgement

In the fore-going sections I have tried to identify what seem to me to be some of the sources of conflict that lie inside our every changing world of assessment. The list has arisen partly from my experience and partly from looking up 'assessment' in dictionaries and thesauri. Etymologically, it seems that the history of assessment is a long one, appearing in the Roman (Latin) term 'assessare', meaning 'to sit down beside'. This gradually morphed into assessment as 'to assist in the office of a judge' and we still have the old word 'assize' as evidence of this ancient strand of assessment. By 1800 it had been formalised into the notion of a judge's assistant (who sat beside the judge) and who was responsible for assessing the value of property for the purpose of taxing it or setting fines. And by 1900 assessment had generalised yet further to the idea of broadly judging the value of a person or idea.

What interests me in this story is that the original meaning of 'sitting down beside' carries the essence of what I believe is good about assessment. Its about spending time (sitting beside the learner) so as *to understanding what has been done* in a piece of work. It describes quite well the viva-voce model of assessment.

In any event, the bottom line with assessment is that *judgements* are made. So it is on that process of judgement-making that I would now like to focus some attention.

The process of making judgements

I have shown in some of my examples that the conventional wisdom in the history of our examination systems (11+, GCE, CSE etc.) leant heavily on the notion of norm-referencing. Standards were defined by the proportion of children in the various groups. The children getting to Grammar schools in Kent in the 1960s were not these that scored X in the test. Rather they were those that were in the top 20% of the cohort.

Criticisms of this norm-referencing approach naturally focused on the hard truth that this it does not identify what pupils can do, but rather measures pupils against each other.

By contrast and in more recent years, GCSE and latterly national curriculum and GNVQ examinations are criterion-referenced. Criterion-referencing was increasingly seen as more educational because it identifies what pupils should be able to do and measures pupils against these identified qualities (not against each other). Indeed in 1985 when I wrote the orange guide for the introduction of the GCSE examinations in CDT I explicitly made this point.

> ...this form is criterion-referenced, and the performance required to achieve a particular mark is therefore specified in advance in the list of criteria on the form...
> (Kimbell for SEC 1986 p 38)

I was younger then but I remain embarrassed by my naivety. In mitigation, I can claim that there was some truth in what I said and I did believe it. But it was only ever a half-truth. And with the benefit of 20:20 hindsight I can now see that there was a gaping flaw embedded in that half-truth. Put bluntly, it is just not possible to define criteria so that they have an exact meaning. This is not because words are not clear but because any intended meaning by the writer is compromised by the personal experience, position and values of the person making the judgements.

The best example I can recall was in the years running up to the implementation of the national curriculum, when a great deal of time and energy was spent on defining excellence criteria by the various NC Working Groups, who (in each subject) sought to define Statements of Attainment (assessment criteria) for ten levels of performance. The following story comes from a very reliable source and recalls the process undertaken in the (1988) Design and Technology (Parkes) group.

It was agreed that a good starting point for defining these criteria (SoA) would be to refine a clear statement for level ten, the ultimate descriptor of what we might expect the most able design and technologists to achieve. The argument ran that if we had such a clear and highly polished statement of capability it might then be possible to work up towards it incrementally; drafting statements that move carefully and incrementally towards this Holy Grail.

So it was drafted, and debated, and redrafted and debated, and edited, and debated and finally it was honed with infinite precision. The group were happy with it as a statement describing the excellence that should be characterised as level ten. And then they showed it to teachers. And the *primary* teachers said "Yes that's what my children do".

There are countless other examples of the same problem. Indeed the technology Order (DES 1990) had 149 of them; and we all grew extremely sick of them.

So what is going wrong?

In this case, as in all others where criterion statements are attempting to define a particular reality, the receiver 'reads' the criterion in a way that makes sense to them. They personalise it. They can't do anything else. And thus arises 90% of the measurement error that attaches to such judgements. Note the story of the hot and cold room. Laming, in his book 'Human Judgement: the eye of the beholder' explains this phenomenon.

> When someone comes to make a judgement…the point of reference is most often taken from past experience. Different people have different accumulations of past experience and for that reason make different judgements about the same issue. We call that difference a 'point of view'
>
> (Laming, 2004 p18)

And having reviewed and rationalised countless examples of human judgement, his startling conclusion to the book is that…"there is no absolute judgment. All judgments are comparisons of one thing with another".

What Laming is saying is that we (teachers) are good at making relative judgements. And we know this to be true. Any teacher will identify for you their 'best scientist', or their 'most musical' child, or their 'weakest reader'. The teacher here is judging against the yardstick of the other children and can make direct comparisons of one with another. These judgements have been shown time and again to be very accurate. Because the personal standards of the teacher have been eliminated as a variable. The yardstick is not a vague criterion in my head but a direct comparison of this child with that child.

Norms, criteria and judgement

The unsettling truth here is that for affective assessment it is necessary for judgement criteria to operate alongside normative standards. Criteria operate in a very different way to norms – and they are both important.

It is as though the criteria set the targets for attention. As an example

- we are interested in reading ability;
- and making sense of the text;
- and being able to identify and interpret mood in the text.

All this is to attempt to refine some *criteria* that can operate as targets for assessment. If I was then to try to score 20 pieces of work on a scale of one to ten, I would start by having a go at locating some on the scale but very soon I find myself cross checking them. If Susan has scored six for that, then Jane must be more…say eight. Until I find one that is much better than both Jane and Susan and that forces me to push them back from six and eight to five and seven. Gradually I work from the criteria whilst at the same time checking scores against each other (this one with that one) to make sure I am happy with the rank. In other words I am using the standards of the group as a *normative* guide.

The harsh reality is that good assessment requires both norms and criteria and our research in TERU has frequently turned on the challenge of understanding how they can best be made to work together. In the two major assessment projects that I have been responsible for (APU in the 1980s and e-scape in the naughties) we have explored exactly these issues.

APU assessment

In the APU final report we outlined a process of assessment that we described as 'fingerprinting' the scripts.

> Our experience of assessment led us to the conviction that it is often easier to identify a high quality piece of design work than it is to say in detail why it is high quality. ...It is interesting to note that in the final analysis our markers were able to make these holistic judgements of excellence at a level of reliability that was significantly higher than that achieved for the assessment of individual aspects of capability.
>
> [But]…It is no good saying "this is good but I don't know why..."...we needed to be able to say why good work was good and what might be done to mediocre work to improve performance. We had to find a way of getting inside the holistic mark into the central traits of good (and poor) performance, but…without defining these traits too rigidly in advance.
>
> To underpin the holistic marking we pursued a complex detective exercise designed to tease out the important qualities of performance from the unimportant or the merely peripheral…We were quite prepared for this exercise to modify how we defined capability.
> (Kimbell et al 1991 p31)

Essentially we took pairs of scripts: a high scoring one (a) and a low scoring one (b).

> We listed all the things that (a) did and (b) did not do, and all the things that (b) did that (a) did not. We therefore identified discriminators of performance from the work itself. We coined the expression 'fingerprinting' the scripts because like a fingerprint each script was unique but by building up a list of discriminating 'yes's

> and 'no's it became possible to describe that uniqueness…Moreover it became possible to ask the computer to generalise these descriptors by selecting all high scorers and printing out their characteristics…While the holistic mark enables us to value a piece of work, the response profiles provide us with a composite description of it.
>
> (Kimbell et al 1991 p32)

APU was not about using an existing set of criteria to measure national performance trends. It was about *trying to find* those criteria that make a difference. So we drew them from the work itself. "Good work is work that has these characteristics". Subsequently in chapters 11-15 of the report we provided extensive exemplification of the levels of capability we have unearthed in the scripts and we reported it in terms of national trends, linked to student general ability, gender and other variables.

So, in a nutshell, what we sought to do was to understand the performances that emerged from the 20,000 APU test scripts and thereby to articulate a view of capability in design and technology: what it looks like and consists of.

E-scape assessment

In the e-scape project we enabled learners to construct web-portfolios of performance in D&T, science and geography. At the point of assessment, we had approx 350 in D&T and 60 each in science and geography. We have reported elsewhere the arguments that led us to a Thurstone pairs model of assessment, suffice it to say here that the digital form of the portfolios made it possible for the first time to develop Thurstone's theory into a working digital prototype system. We christened it the 'pairs engine'. We optimised the portfolio display so that whole portfolios could be viewed (in thumb-nail format) on 20 inch monitors. This portfolio-at-a-glance arrangement allowed judges to scan the whole piece of work. Additionally however, judges could click on any of the elements in the boxes (e.g. photos/drawings/sound files) that would then automatically jumped to full screen images or play as voice/image/video files.

> The pairs engine managed the assessment process. The system is based on a theory initially developed by Thurstone (1927) concerning the reliability of judgements. This theory was developed by Pollitt (2004) for inter-board reliability studies for GCSE and other school-based examinations; checking the reliability of the assessments that had been made. For phase three of e-scape we then developed the system further so that the pairs judgements became the front-line assessments of the portfolios.
>
> The 'pairs engine' presents a judge with pairs of portfolios and the judge has to scrutinise the work and make a balancing holistic judgement about which of the portfolios represents the greater capability. For the design and technology sample we had 350 portfolios and 28 judges, each of whom made 130 paired comparisons. The geography and science samples were smaller and had judging teams of six.
>
> The judgement process is based on criteria, but these are not scored directly – but rather are interpreted by the judge into a single holistic judgement. At the outset the engine assumes that all the portfolios are of equal quality, so judges might well be presented with work that is radically different in quality. These judgments are easy and quick. As the data begins to build however, the engine begins to estimate a rank order and thereby presents judges with portfolios that are closer in quality. These judgements are more difficult and require the judge to look deeper into the portfolios to identify discriminating features.
>
> Eventually a complete rank order emerges – and with very high inter-judge reliability. For each portfolio the engine generates a 'misfit' statistic – essentially reflecting the amount of disagreement between judges that it created. Moreover, for each judge the engine generates a misfit statistic – reflecting the consensuality of that judge with the rest of the judging team. If either misfit statistic goes above an acceptable level, remedial actions are triggered. The remarkably high reliability of the judgement process (0.95) is explained by the fact that each portfolio is compared to approximately 20 others and is seen by more than 20 judges. What emerges from the pairs engine is in effect the professional consensus of the expert team of judges. The same levels of reliability were achieved with the science and the geography judging teams looking at their portfolios.
>
> (Kimbell et al 2009 p 56)

I have been very aware, since I first saw the Thurstone pairs model of assessment, that it is something very, very different. We have reflected endlessly on it within the research team for we are aware that there is the potential for the pairs engine to stand assessment on its head. The priorities and practices of the known world of assessment have been changed. And we have speculated on what this new world might look like. It is appropriate to think of the change in the context of the transition from the modern to the post-modern for there are at least three features of the pairs-engine process that fit very easily into a post-modern debate.

RESEARCH

Post-modern assessment

A post-modern culture might be characterised as one that rejects objective truth and sharp classifications or hierarchies of black/white, male/female, straight/gay. It stands for a rejection of modernism's scientific mentality. Whereas modernism was associated with authority, identity and certainty (science as objective truth) postmodernism is more associated with difference, separation, scepticism (science as a provisional consensus).

The assessment world, one might suppose, is deeply rooted in the modernist tradition. It assumes that there are 'those in authority' who know about good and poor and can therefore appropriately pass judgement on others' work. So awarding bodies have examining teams and these teams have senior examiners. All very hierarchical. How else does assessment make any sense? Moreover, the idea that there is a 'right' way of doing designing that can somehow be 'added-up' is a deeply modernist form of thinking. By contrast, within a post-modern critique, everyone is a designer; there are lots of ways of being a designer; and many different ways of being good as a designer. Interestingly we have always argued for this position.

So lets look at assessment through the lens of this post-modern critique.

Firstly, it relies on a democratic view of assessment in which multiple judges contribute to the overall result. This is NOT a top-down hierarchy with a senior examiner deciding standards. Rather it is a very flat structure in which your judgements are as good as mine so long as the misfit statistic does not get too big. This misfit statistic is a reflection of the *power of the collective*, and the standard emerges from the consensus of all the judging team. Every teacher in the e-scape trials was a judge and was therefore contributing to the establishment of the standard.

Secondly, just as with APU the judging process starts with holistic judgements. This is not to say however that it ends there. Having created the rank order (from the engine) it is then necessary to tease apart what has been going in *inside* the judgements to establish what it is that acted as discriminating factors in judges' decision-making. This we have done in the e-scape phase three final report. In chapter 11 of that report (pages 78-134) we characterise the types of performance that are evident in the work. We have long held the view that there is more than one way to be good at designing, and through this diversity of portfolio data we are able to illustrate several of the ways in which portfolios have been judged as good/OK/poor.

Thirdly, taking the democratic notion a step further, there is a perfectly valid question of who will be the judges, and in this case we modelled what would happen in the learners themselves became the judges. Placing them at the receiving end of the pairs engine reveals all sorts of things about how they (as students) value their portfolios. Perhaps not surprisingly their rank order (for a selected sub-sample) was exactly the same as that generated by the 'real' judging team. In terms of modernist reliability statistics this is fascinating but in terms of a post-modern re-ordering of the nature and purpose of assessment it is perhaps more significant. The students' reaction to the process can be paraphrased as "why didn't you show us this before we did the activity" (i.e. *before* they constructed their portfolios). Because..."I would have told my story differently and clearer".

The e-scape portfolios are essentially real-time narratives, created by the learner, as he/she works through a process of designing a prototype solution. These narratives are illuminated through multiple media (text/image/voice/video) and enable learners to construct a very personal account of their journey. Thereafter, the process of engaging with the portfolios through the pairs engine enables the originator and the judge to review the narrative on many levels. In learning terms the discussion might be about how such narratives might be enriched. But in assessment terms the debate is about how they might be valued.

We have speculated on a new world order (for assessment) in which all student work within an examination goes into a single national pot, and is put through the pairs engine with every teacher who enters students for that examination acting as a judge. Assuming (on average) perhaps a 25:1 ratio (25 learners for every teacher), there might be 25,000 learners and 1000 judges/teachers. We have modelled the numbers and the time taken for this processs, and it would be quicker and simpler than the current arrangements. The big changes however would be (i) that the resulting rank-order would reflect the consensus of those teachers, and (ii) in the process of generating it the teachers would have seen a real cross section of the whole national sample. The teachers would be doubly empowered by the process. It would also be simple to pause the process at any point and ask teachers to do some of the judgements acting as a small team. The point of this process would be to promote teacher-discourse (on-line) about *why* this one is better than that one. The role of the exam board would be merely to decide grade boundary positions across the rank, and even that could be done statistically.

Perversely, what looks at first sight like a mechanistic, software-driven process, turns out to be far more humane than the current arrangements. Teachers have the ultimate control over judgements and their holistic judgements can allow for many different approaches to designing (see chapter 11 of the e-scape report). On the other side of the coin (viewed from the exam board) the hard statistics are unarguable. Judgements are far more reliable than in the current arrangements; appeals from schools would have no meaning, since the final position is a consensus of multiple judges; and the thorny problem of trends over time (so beloved of politicians) also goes away since each year's pot of portfolios can have 'marker' portfolios added that are taken from previous year's grade boundaries. Where these markers emerge in the new rank indicates (magically) last year's boundaries within this year's rank.

Viewed through this lens, the challenge of the *transient* and the *timeless* looks somewhat different. Our e-scape portfolios, being web-based, might seem unusually ephemeral and therefore transient. And yet they open up the possibility of solving some of the most timeless and intractable problems of educational assessment.

References

Blair T, 1999 Speech at the Millennium Products Awards – Millennium Dome London Tues 14th Dec 1999
(see www.design-council.org.uk)

Craft A, 1997 Can you teach creativity? *Education Now* Publishing Co-operative. London

Haddon F A & Lytton H, 1968 "Teaching approaches and the development of divergent thinking abilities in primary schools" *British Journal of Educational Psychology* vol 38 1968 pp171-80

Kimbell R, 1986 *Craft Design & Technology A guide for teachers*, Open University Press for SEC

Kimbell R, Stables K, Wheeler T, Wozniak A, and Kelly V, 1991 The Assessment of Performance in Design & Technology London School Examinations and Assessment Council/Central Office of Information

Kimbell R, Wheeler T, Stables K, Shepard, Pollitt A, Whitehouse, Martin, Lambert, Davies. 2009 *E-scape portfolio assessment phase 3 report TERU:* Goldsmiths University of London

Laming, D. (2004) *Human Judgment: the eye of the beholder.* London, Thomson.

Pollitt A & Ahmed A, 2009 The importance of being valid A paper presented at the *10th Annual Conference of the Association for Educational Assessment* – Europe. Cambridge Exam Research www.camexam.co.uk Malta, November 2009

Pollitt A, (2004). "Let's stop marking exams". Paper given at the *IAEA Conference*, Philadelphia, September. Available at:
http://www.cambridgeassessment.org.uk/research/confproceedingsetc/IAEA2004AP

Thurstone, LL. (1927) A law of Comparative judgement. *Psychological Review,* 34, 273-286

r.kimbell@gold.ac.uk