e-assessment in project e-scape

Professor Richard Kimbell, Technology Education Research Unit, Goldsmiths, University of London

RESEARCH

Abstract
In this paper I shall present a new model of assessment – derived specifically in the context of the latest research and development project being undertaken in TERU at Goldsmiths College. The paper starts with an analysis of some of the roots of our assessment practice in norm and criterion referencing, and I indicate some of the pitfalls of the criterion referencing approach that was developed as part of the England/Wales National Curriculum in the early 1990s. I then describe what we have done over the last couple of years in TERU to enable learners to create e-portfolios for assessment within project e-scape. Using these e-portfolios, I then present a completely new approach to assessment developed in association with the former head of research at the University of Cambridge Local Examinations Syndicate. The approach was (on the face of it) simple, using a web-based interface. But the outcome was astonishing – producing reliability statistics of 0.93: far higher than can be achieved in normal coursework assessment. The issues that this approach raises are many and profound, not the least of which is the possibility of a new, trustworthy, conceptualisation of what coursework assessment might become.
I conclude with a brief projection into the future – both in terms of how e-scape is planned to develop over the next few years and in terms of what its consequences might be in the classroom.

Key words
project e-scape, assessment, TERU, e-portfolios

Norms and criteria

I have always been a bit suspicious of criterion-based assessment.

It is not that criteria are not important in any assessment regime, because of course they are. It is just that, rather like recent converts to a new faith are likely to be the most ardent, the 'conversion' process I was witness to in the UK from the early 1980s (from norm-referenced to criterion-referenced assessment) tended to carry with it a highly simplistic formula:

> Norm-referencing is bad and even immoral (it compares children with each other)
>
> Criterion-referencing is good and morally preferable (it is based on objective external truths).

Inevitably, the harsh reality of assessment in schools soon demonstrated that the two are not mutually exclusive, but – rather – are dependant upon each other.

In the early 1980s I recall that the developments that led eventually to GCSE were focused originally on a fantastically detailed analysis of 'grade-related-criteria'. From that moment onwards learners were no longer to be judged against the norms of their class/group/cohort. Rather they were to be judged against absolute statements of capability. The concepts of 'better than' or 'worse than' were inapplicable – and even unclean. From here on we would have positive statements of what learners know, understand, and can do, set at the many and various levels of capability for which the assessment system was designed. It seemed more equitable, more thorough, and (as I have suggested above) more morally desireable. But even then there seemed to me to be something a bit simplistic about it. It made me nervous… but I could not quite see why.

It took another five years or so for me to be convinced that it was a dangerous delusion to believe that criterion-based assessment was a judgement process based on free-standing truth. By that time we were up to our eyes in criteria – only they were called "Statements of Attainment" (SoA). These were the cornerstones of National Curriculum Assessment, perhaps the most wasteful and destructive experiment in assessment that the world has ever seen. SEAC – the School Examinations and Assessment Council – was in total control, and those of us who were involved in trying to evolve meaningful and helpful assessment activities were driven before the storm of SoA. We should remember that there were approximately 150 such Statements for design and technology, distributed across 10 levels and several Attainment Targets – ATs – (initially 5, then 4, then 2, and now 1).

The process of assessment was reduced to box-ticking. Can they do… this, this, this, and that? If so they are Level 3 for AT1. Now we'll check more for AT2, and yet more for AT3, and even more for AT4. Then we have to aggregate all these positive 'can-do' statements and arrive at an "answer" for the supposed level of attainment of the learner. With 150 SoA, and (say) 20 learners in a class, that's a mere 3,000 boxes to tick (or cross). Teachers were driven mad by the process, and (as night follows day) the sad Secretary of State for Education paid the price. He was sacked for being in charge when (in 1992/3) teachers finally flexed their muscles and imposed a national boycott of the tests.

The boycott was brought about by many factors, but underlying teachers' discontent were at least three key issues:

Proliferation: It is almost inevitably true that if one seeks to define the whole of D&T capability in a set of Statements – they either have to be very generalised or very numerous or both. Those for the England/ Wales NC managed to be both – simultaneously.

Prescription: It's true that *proliferation* was a problem, but the associated problem was that the Statements (if they meant anything at all) necessarily meant that progress was defined by doing Level 1 things before Level 2; and those before Level 3 and so on. So they carried the assumption that there is a *right way* to be excellent (or poor or just OK). Experience however suggests that learners can be excellent in design and technology in dramatically *different ways*. The tendency was for SoA *descriptions* of excellence to transform insidiously in *prescriptions* of that excellence. "Do it my way – it's the only way that counts."

Meaninglessness: But even that wasn't the real core problem, which was that the SoA tended – on their own – to be *meaningless*. Let me give you a couple of examples. They are a couple of the SoA taken from the levels of the 1990 version of the NC for England and Wales.
Do learners… "*use specialist modelling techniques to develop design proposals*" Yes or No?
What level of capability is this statement seeking to encapsulate? Is it Level 2 (e.g. Plasticene or Duplo as a specialist modelling technique), or perhaps Level 4 (e.g. LEGO Technic as a specialist modelling technique), or perhaps Level 6 (eg. Crocodile Clips as a specialist modelling technique), or even Level 10 (e.g. Pro/ENGINEER CAD as a specialist modelling technique).

Or how about this one… do they… "*use drawings and modelling including annotated drawings and working models to develop their design proposals?*" Again, **what kind** of working model? A CAD animation? A LEGO Technic wheel/axle? A pop-up folded card?

The reality of such statements was that they utterly failed to capture the level of capability that they sought to describe, because the reader had to ***interpret*** the criteria so that it meant something concrete. Ironically therefore, the criteria only meant anything when you knew the level they were written for. If I tell you that the first one was Level 6 and the latter one was Level 3 then they start to take on some meaning. Aha you say… Level 6… that means KS3/4 and so the '*specialist modelling technique*' can't mean plasticene…but it might mean Crocodile Clips. The criterion, far from ***replacing*** a norm, only acquired meaning when a norm was imposed upon it.

That is why SEAC spent endless millions of pounds on a process called 'exemplification'. In order for their criteria of capability to acquire meaning for teachers, SEAC published book after book after book of learners' work exemplifying the levels at which performance was to be assessed. What they were trying to do was to define the meaningless statements by reference to real meaningful work. "…This is what we mean by this statement at this level. And that is how it is different for the next level."

If I am trying to make a judgement about this piece of work, and all I have to measure it against is an abstract criterion, it proves quite impossible. But as soon as I have some exemplification of it (in the form of another child's piece of work) …then aha… *that's* what you mean by that criterion at that level… and its obvious that the piece I am trying to mark is better than that. The criterion – on it's own – does not help. The *comparator* makes sense of it.

RESEARCH

Accordingly, we should recognise that the moral high ground of criterion-referencing is not quite as moral as we first thought, because it is still dependent on comparing one child's work with another. It is just that our norm-referencing is once-removed. We had to normalise the *criteria* to make them meaningful.

As Laming so perceptively argues in his recent book on the psychology of judgement.

> *"There is no absolute judgment. All judgments are comparisons of one thing with another."*
>
> (Laming, 2004)

All this is by way of introduction to a somewhat scary new world of assessment to which I have recently been introduced and that I wish I had known about twenty years ago. In order to make sense of the story however I need first to take you on a brief detour into our most recent research venture in TERU.

Project e-scape

In 2003 we were invited by DfES and QCA to develop a new approach to assessment for GCSE that might better reward learners' *innovative* performance and *teamwork*. Over a two year period, in association with the Awarding Bodies, we worked up the concept of a six-hour structured activity (two consecutive three-hour mornings) in which learners take a design task from its starting point up to the point of a working prototype. We prioritised *concept development* and *modelling*, working on a folding-out A2 worksheet, drawing on collaborative techniques that both enriched learners' starting points and helped them to keep their ideas on track with reflective critiques. The outcome of the national trials in June/July 04 were so encouraging that one of the English GCSE Examination Awarding Bodies (OCR) decided to develop our prototype as part of their new Product Design specification. Their excellent 'Innovation Challenge' (http://www.ocr.org.uk/Data/publications/teacher_support_and_coursework_guidance/GCSE_Desig59383.pdf) is the current manifestation of the research we undertook through 2003/4.

It was while we were developing the six hour task that we began to explore the possibilities of digital tools to enhance the activity and we took a new proposal to QCA. The outcome was ***project e-scape***, (Kimbell et.al., 2007) in which we have sought to create a digital version of the six-hour activity. In essence this involves finding ways to replace the A2 worksheet with an e-portfolio in a web-space. The clever bit of this project (at the classroom end) lies in the fact that the e-portfolio is unlike anything that currently exists by that name. Typically such things are second hand re-constructions of real designing – in PowerPoint (PP) or some other sequential software.
The construction of the e-portfolio is typically a different task to the designing that it seeks to illustrate. First do your designing – then tell the story in your PP e-portfolio. By contrast the **e-scape** system uses hand-held digital tools directly in the nitty-gritty of the designing activity in workshops and studios. As learners do their thing, the hand-held digital tools up-link the work dynamically into a secure web-space, where their e-portfolios emerge before their eyes as they work through the activity. These are *real-time* design e-portfolios.

Having explored a number of tools as the basis of this system (digital pens was our first experiment and preference), we settled on Personal Digital Assistants (PDAs) for the national pilot. The strength of PDAs is their multi-tasking; enabling learners to use them like digital sketchbooks, digital notebooks, digital cameras and digital voice-recorders. All these data types have been integrated into the e-scape interface, and the emerging portfolios are rich multi-media, real-time accounts of learners' six-hour struggles with their design task.

The national pilot of this system took place June/July 2006 in 14 English schools from Cornwall to the Scottish border. The learner groups lapped it up – it fits very smoothly into their mobile computing youth culture. Teachers were more apprehensive, but by the second three-hour morning were invariably keen to explore the system and discuss its wider applicability.

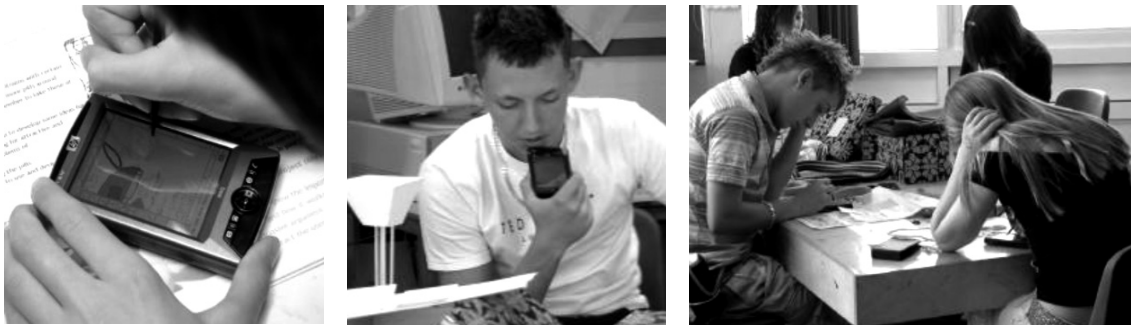
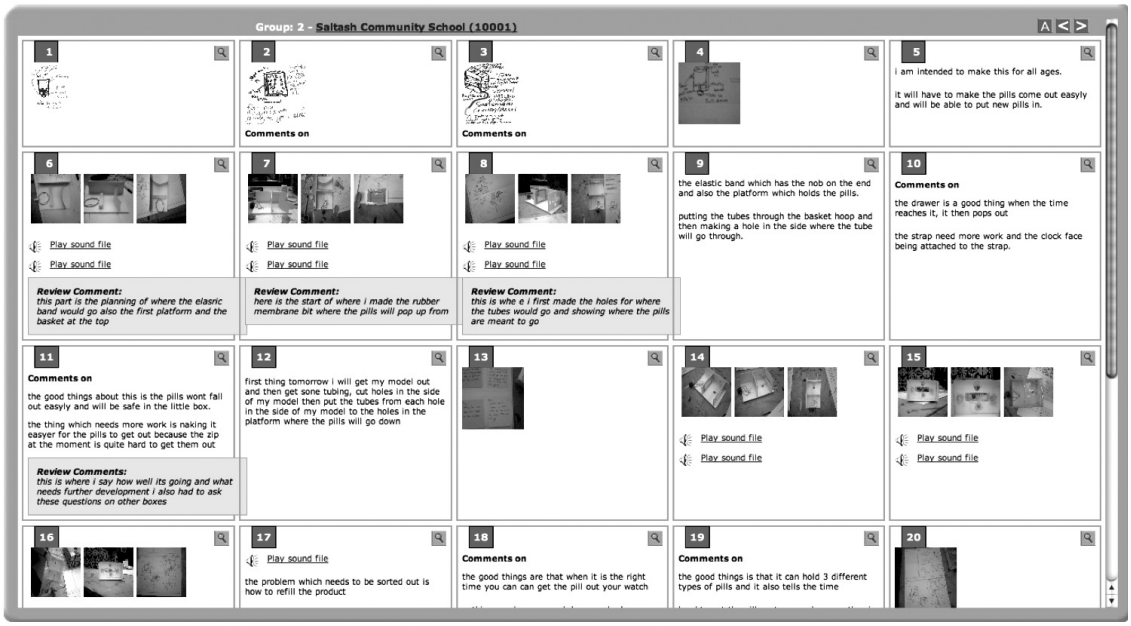


Figure 1. Hand-held digital tools leave a dynamic evidence-trail e-portfolio in the web-space



While we were developing the system, Teachers' TV made several programmes about e-scape and they can be found at the following URLs, free to download.

The future's handheld (an account of one of our two-day assessment trials in a school in Cornwall).
http://www.teachers.tv/video/3306
First broadcast Monday 8th January 2007 – 5pm

New Technology – the issues (talking heads analysing the two-day trial).
http://www.teachers.tv/video/3307
First broadcast Monday 8th January 2007 – 5.15pm

e-assessment-where next? (broader e-assessment policy, based on e-scape innovation).
http://www.teachers.tv/video/5431
First broadcast Tues 9th January 2007 – 12.30pm

RESEARCH

The successful conclusion of phase 2 of project e-scape raised many issues of importance for the future of e-learning and e-assessment.

Concerning **technological** challenges, the whole system is driven by a remote server dynamically sending and receiving data to and from hand-held digital tools, putting the teacher in control of the sequences of the task and automatically building an evidence trail in the web portfolio.

Concerning **pedagogic** challenges, everything we did for the purposes of collecting evidence for assessment **also** helped to scaffold the progress of the activity and the performance of learners.

Concerning the **manageability** challenges, the key point is the infusion of technology into activity. Real-time activity in studios, workshops, playing fields, theatres, science labs and the like, is typically not aligned with digital power. That power typically sits in splendid isolation in the shimmering purity of ICT suites. In e-scape we have shown how the technology can get down and dirty and unleash its digital power where it is really needed. And in the national pilot we demonstrated that it was manageable.

To this extent, e-scape has been a success, and the evidence is in the website which contains approx 300 e-portfolios full of integrated drawings, notes, photos and the real authentic voice of the learners explaining what they are doing and why. But, arguably, this is not what e-scape will be remembered for. It seems likely that it will be remembered more as the first occasion on which a completely new mechanism of assessment has been used. It is not that e-portfolios are new, it is rather that having all these e-portfolios in the website enabled us to launch into a different model of assessment. And to present the argument, I have to take us back to Laming and the problem of making judgements.

Assessment within e-scape

Just as with the previous project, we have conducted e-scape in close association with the GCSE Awarding Bodies, and it was through this route that we were introduced to a scary individual who tends to turn upside-down all the normal preconceptions about assessment and whose ambition (he wants it carved on his tombstone) is to be the man who got rid of marking. Alistair Pollitt – who was at one time the director of assessment research at the University of Cambridge Local Examinations Syndicate – drew our attention to a system of assessment that was first articulated by Thurstone in the 1920s.

> The alternative approach to summative assessment that I would like to propose is based on the psychophysical research of Louis L. Thurstone, and specifically on his *Law of Comparative Judgement* (Thurstone, 1927)… The essential point will be familiar to anyone grounded in the principles of Rasch models: when a judge compares two performances (using their own personal 'standard' or internalised criteria) **the judge's standard cancels out**... a similar effect occurs in sport: when two contestants or teams meet, the 'better' team is likely to win, whatever the absolute standard of the competition and irrespective of the expectations of any judge who might be involved.
> (Pollitt 2004 p6)

Currently, for GCSE awards (age 16) in England and Wales, assessment is undertaken by drawing up a list of criteria for the performance; allocating a block of marks to each criterion; and judging individual pieces of work to decide how many marks to award for that criterion. This involves a judgement of the individual piece of work against the criteria and the problem that I described earlier still remains. Am I to award 6/15 for this piece on this criterion, or 7/15... or 8/15? What does 7/15 mean? Well, go back to the exemplars and we have one there as a guide.

Interestingly there are two circumstances in which 7/15 can be seen to mean something. First, when we have an exemplar of it, we can compare our piece to be assessed directly with the exemplar. But what if the exemplars are not at every level? What if (more likely) we have four exemplars distributed across the scale, e.g. for scoring 13+/15, 10-12/15, 5-9/15, and <5/15? Then we are forced to retain the concepts of 'better than' and 'worse than' to use in association with the exemplars. Yet more watering-down of the moral high ground.

Pollitt recognised the inevitability that assessment requires the comparison of work from one learner

with work from another. And, if all assessment is really about comparing one piece of work with another, why don't we just compare them directly?

Based on this idea, he proposed a system in which judges compare two portfolios and decide merely which of the two is the better. The judges of course have to have some notion of what might be meant by 'better' and 'worse', so some shared values are important and these would helpfully be articulated as a set of criteria. But more of that later. They key point here is that criteria are not 'marked' as they conventionally are. Rather, a holistic judgement is made about which piece of work – overall – best represents an excellent piece of work. One of the beauties of this (Thurstone) model is that the idiosyncratic standards of the judges just don't matter. I may be a hard marker or a soft one – but I still have to decide which of the two pieces is the better. Judges' personal standards (the greatest source of error in current assessment procedures for 16+ GCSE exams) therefore just cancel out.

The greater the true difference between the quality of the two portfolios that I am examining, the more likely it is that the better one will win each time they are compared. Thus a large set of comparisons does more than just generate a rank order; the relative frequency of success of one performance against another also indicates how far apart they are in quality.

> *Statistical analysis of a matrix of comparative judgements of 'scripts' can construct a measurement scale expressing the relative value of the performances. The result of comparisons of this kind is* ***objective relative measurement****, on a scale with a constant unit. Furthermore, if a few scripts that have already been agreed to represent grade boundaries – perhaps from a previous sitting of the examination – are included in the comparisons, the whole process of marking, grading and comparability of standards can be replaced by the collection and analysis of paired comparative judgements. (Pollitt 2006 p2)* ***(my emphasis)***

In a trial assessment – based on a set of work that we had formerly marked through conventional (number-based) procedures – we showed that there is a strong relationship between the parameters derived from the comparative pairs judging and the marks previously awarded. The value of R2 was 0.81, corresponding to a correlation of 0.90 between two linear variables, as high as could be expected in a case like this. Armed with such strong evidence, we felt confident to embark on the assessment of the whole *e-scape* sample using the Thurstone/Pollitt approach of comparative pairs judgement.

The judging process for the main body of e-scape data was designed in three rounds, using seven judges. Round one involved each of the judges reviewing 140 pairs and deciding on the 'winner' in each case. The general response of judges was that the early pairs (say the first 20 or 30) took as long as 10 minutes per pair to decide, but gradually we got quicker. This speeding-up resulted in part from being more skilled in working our way around the web-based portfolio, and in part from the fact that the pairings inevitably threw up repeats. Having got properly inside a piece of work at the first time of asking, it took only a much briefer scan second time around to remind us of its qualities. By the end of the 140 pairs we were typically doing each pair in two minutes.

Having completed round one judging, Pollitt analysed the results into a rank order and round two pairings were then selected to refine and confirm the order. The pairs this time tended to be closer together, comparing (for example) something like a B with a C, or a D with an E, whereas round one (being random) had just as frequently required us to compare an A with an E. The relative difficulty of these round two decisions was offset by the familiarity (by now) with much of the work. Generally round two was quicker than round one. Round three of the judging was then focused on enriching the data (so at to assure reliability) at the grade boundaries (A/B/C/D/E).

RESEARCH

The complete set of results can best be seen in the chart here (Figure 2). In it, the portfolios have been sorted into order and are shown with their standard errors. Pollitt reported to us as follows:

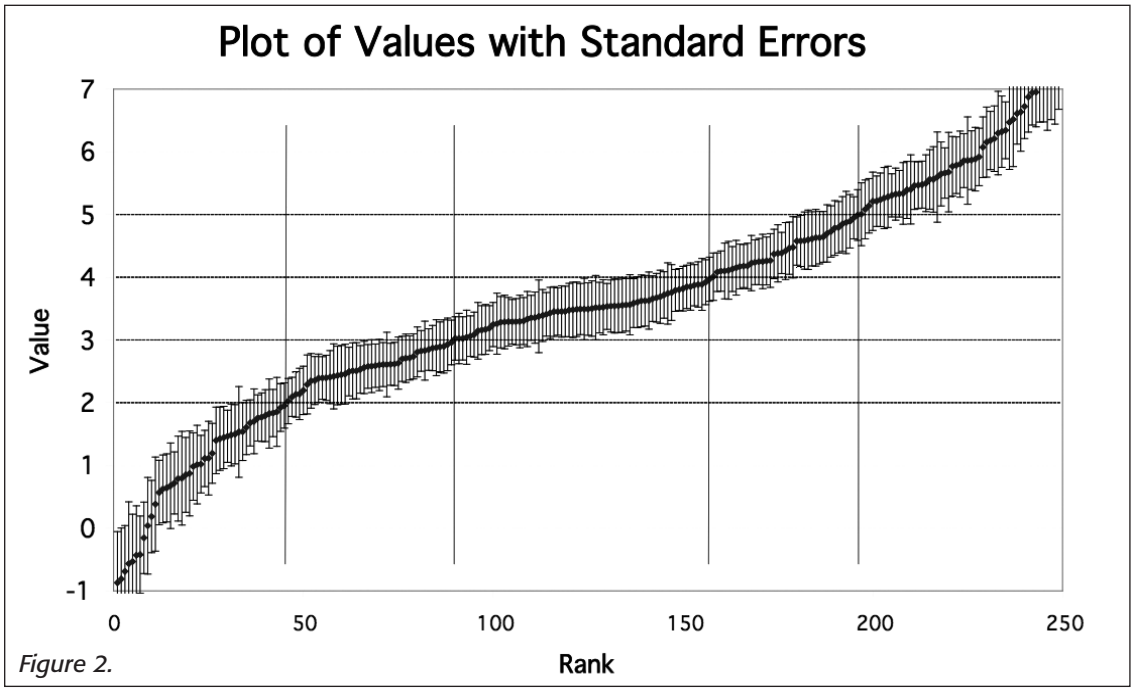


Figure 2.

In formal statistical terms 68% of the portfolios' "true" values will lie within one standard error of the reported value. Vertical lines are drawn through the five grade boundaries (A-E) to show how many pupils would fall into each grade. The analysis of the judgements also gives a traditional indication of the quality of the measurement process and the key figure here is the reliability coefficient of 0.93. This figure allows for unreliability between markers *as well as* for lack of internal consistency within the examination – most traditional reliability coefficients only allow for one of these. Only a few current GCSEs are likely to be as reliable as this if we consider both sources of unreliability.

(Pollitt 2006 p5)

That then is the bare bones of the system and the approach we used to operationalise it as part of the e-scape research project. Whilst we undertook the exercise in association with Awarding Bodies, we were not able to work right through the awarding process that would have required further levels of analysis and judging. We did however simulate the process of identifying five notional grades (A-E) and explored the means for assuring reliability at those grade boundaries. The whole of this has been fully reported to DfES and QCA, (see Kimbell et al 2007) and in that report we identify three nuggets of information to which I would – in particular – draw the attention of readers, quite apart from the performance scale itself.

First the *reliability* of the resulting scale.

"The key figure here is the reliability coefficient of 0.93. This figure allows for unreliability between markers as well as for lack of internal consistency within the examination – most traditional reliability coefficients only allow for one of these. Only a few current GCSEs are likely to be as reliable as this if we consider both sources of unreliability."

But this reliability is hardly surprising. Each piece of work has been compared with many others, (17 as a minimum) and the judgments had been made by many judges. Any idiosyncratic judgments were soon outweighed by the weight of opinion of the

team. The process is almost inevitably more reliable than current GCSE practices, where much of the work is assessed by the teacher alone, or at best by the teacher and one external moderator.

Second it is important to note the consistency of the judges. In this comparative pairs approach, the analysis automatically produces a reading of the judging team, specifically concerning their consensuality. The system notes how often – and by how much – my judgments are at variance with the other judges and in the end produces a mean score for the whole sample. If I am more than two Standard Deviations from that score, then I am a cause for concern.

"None of these judges fails the test"

Third, the system also automatically produces data on the consensuality of judgments applied to individual portfolios. Reference to the 'plot of values' (above) shows some portfolios with much longer standard error 'tails' than others. These are the portfolios over which there was a considerable amount of disagreement within the judging team. In the process, the system automatically highlights the pieces of work that need closer attention.

*"It shows that a few of them ought to be checked (at least 2 of the 249). The criterion would be 0.85+2*0.23, or 1.31; portfolio number 247 exceeds this, suggesting that there is something about it that is unusual enough to warrant a further look – perhaps different judges valued them in different ways."*

These three key qualities are all automatic virtues of the comparative pairs judging process.

(Kimbell, Wheeler, Miller and Pollitt 2007 p63-4)

Having proved (in e-scape phase 1) that hand-held technologies in the school workshop could link directly to web-portfolios, we were asked to take the concept on to phase 2. In this we have built a working prototype system and operated it successfully in 14 schools across the country. Moreover we have developed the comparative pairs methodology for assessment and shown (for the first time anywhere in the world) that it can work as a means for front-line assessment of learners' e-portfolios.

Coursework assessment

It was whilst we were in the middle of this assessment/judging process, that the political storm broke about the role of coursework in GCSE assessment. In brief, the issue seem to be that coursework components (for high stakes assessment) are just not trustworthy, and for two reasons:

- we cannot be sure whose work it is (parents/ teachers 'helping out'?);
- we cannot ensure high levels of reliability in the assessment process.

Accordingly, for some subjects, the whole concept of coursework is now in dispute, and means have to be found to make it more politically acceptable. It is in this context that e-scape seems to offer at least one intriguing way forward.

It is quite possible to see e-scape as a form of coursework with learners producing their own creative solutions to tasks that are set and administered in a school setting. This solves the first trustworthiness problem outlined above. Then the assessment process – based on 'differentiated pairs' – solves the second trustworthiness problem by producing highly reliable judgement data.

The outcomes of phase 2 of e-scape have been intriguing at a number of levels, but perhaps the most interesting of all is the different light that it throws on the debate about norm/criterion referencing for assessment. Actually there is sadly not much debate on this topic, since – being unclean – norm referencing has not had much of a run in the last few years. All assessment (at least for examination purposes) has been focused on criteria and 'outcomes'.

Norms and criteria revisited

I have always argued (see for example Ch 6 in Kimbell 1997) that this preoccupation with criterion-based assessment was a peculiarly one-eyed approach, particularly since teachers find it really easy to rank-order the learners in their groups. Whatever the difficulty they have in deciding how many marks to give learners in relation to this or that criterion, they have no such difficulty in deciding that Katy is more capable than John. As Laming points out – all judgements are (in the end) comparisons of one thing with another.

RESEARCH

So where does this leave e-scape in any putative norms/criterion debate?

Whilst we did not set out to do this in the first place (our methodology evolved somewhat from the original plan), a retrospective analysis of our approach suggests that whilst the judging process appears to be norm-referenced, there are three key points of interaction with criteria.

First, the task and the activity were designed in relation to the **criteria** that define capability in design and technology. Using that starting point, we sought ways to evidence learners' capability and we gradually evolved the tasks and activity structures that led us through to the national pilot.

Second, the judging process was based on criteria in the sense that judges were asked to 'hold-in-mind' the four key criteria that defined capability within this task. Judges were not of course asked to judge against these qualities individually, but they were asked to hold them in mind as the main qualities we were using to inform the ***holistic*** judgment of capability.

From that point on, the assessment process was entirely **norm-referenced**, judging one learner against another, and another, and another. And the rank-order that emerged from Pollitt's analysis of our judging was just that – a rank from the best performer to the worst. Intriguingly however, there is subsequently an opportunity for the criterion-referencing process to re-appear in our post-judging analysis of this rank for awarding purposes.

Third therefore, whilst we speculated on the existence of five grades (A-E) just to check out the process of enriching data at grade-boundaries, in reality the process of awarding grades is criterion-related. We *could* therefore go back into the ranking, pull up individual pieces and decide (in relation to the criteria) whether this piece of work represents an A/B/C etc. The process could be far easier than currently is the case however, for one could imagine going into the ranking, identifying an individual data point (representing an individual learner), clicking on it, and having that portfolio come up live on screen. Then we could click on adjacent ones up and down the rank to assure ourselves that the A/B boundary is correctly located according to the criterion-based performance of the pieces of work.

So e-scape in reality represents a blend of norm and criterion referencing. Rather than doing both simultaneously however as GCSE markers currently do (even though they are not supposed to), we have separated the processes.

- criterion reference the task, the activity and the mind-set of the judges;
- norm-reference the work;
- criterion reference the awarding.

It's a bit like a weights and measures system. I understand that a 'gram' (as a unit of measurement) is the weight of a cubic centimetre of water. But if I buy 200 grams of cheese in a market in France, they don't decide on how much cheese to give me by going back to the original meaning of the unit and weighing my cheese against 200cc of water. They use simple scales that are appropriately calibrated to represent the reality. So too with our judging. We do not measure each piece of work against the original meaning of the criteria. Rather we use learners' work as a comparative scale. And just as the cheese-seller might be required to demonstrate (every now and then) that her scales are accurate – so too would we (for accurate awarding) reference our scale to performance criteria.

Who should do the judging?

Having evolved the system of pairs judging over the last few months, we have begun to speculate on what might happen if we implemented it more widely. And at the top of the list of interesting questions (relating to scaling it up so that it could become a national system of assessment) is who would do all the judging? We had seven judges for 249 portfolios, but what if we had 2,490, or 24,900?

I'm not sure whether the number of judges would need to expand proportionately – to 70 and 700 respectively. The statisticians will no doubt inform that question. But what would happen if we asked teachers to do the judging? Submitting learners for current GCSE assessment arrangements involves teachers doing the front line assessments. And they

have a pretty tough time around Easter getting on top of all that marking of portfolios. What if they did a slice of judging instead? And suppose that they were presented with some (not all) of their own learner group to be compared with many others from other schools and regions. At a stroke, teachers get a real glimpse of the breadth of styles and standards of work that were formerly in the private playground of Awarding Body examiners. What a potentially valuable professional experience.

But what about reliability? Could we trust them?

As I pointed out above, the judging system automatically generates a consensuality measure for each judge. If, for whatever reason, individual judges are wayward in their decisions, out of line with the mass of judges, then it would soon become apparent. And various alternatives would then be available. Thinking defensively, one could simply remove those judges and re-calibrate the rank using more consensual judges. Or, more positively, one could talk to them and see whether better training is all that is required to bring them into line. But even more interestingly, what if they are making different decisions because they are seeing something important that others do not see. Today's norms for design and technology were created by yesterday's mavericks. So we should not ignore the possibility that the non-consensual judges might be right.

I can imagine a world in which every teacher (not just Awarding Body moderators) is involved in contributing to national standards. It could democratise front-line assessment and leave the Awarding Bodies to monitor the process and subsequently to identify the grade boundaries.

…and the classroom consequence… what of formative assessment?

I am aware that pairs judging might incur some displeasure. Not only is it *normative* (at least on the surface… and at least partly), but also it is *summative*. It is an approach whose main strength would appear to be to provide reliable measures of performance at the end of a programme of study. For most teachers, summative assessment is an unsavoury necessity. *Assessment for learning* is where the heart is. So it might be worth a brief diversion into the classroom consequences that might flow from a 'differentiated pairs' approach to assessment.

Current portfolio practice in design and technology tends to centre on how to get maximum marks for *investigating*, or *generating ideas*, or *evaluating*. All the **bits** of the portfolio that have chunks of marks allocated to them tend to become learning targets. And the consequence has been endlessly extended portfolios that Awarding Bodies have (in the last few years) been forced to limit to a fixed number of pages.

The expansion (and the prettying-up) of paper portfolios was the direct result of assessment of learning. Teachers could see what learners needed to do – so they made sure that learners provided it. And the portfolios got bigger and bigger and bigger. No-one is to blame for this. Teachers were acting in the best interests of their learners, and (mostly) the learners responded. But the effect was for portfolios to grow to the point of unmanageability, and eventually the Awarding Bodies had to act to limit things. They set a page limit on the portfolios. But the subsequent effect of this (through the same mechanism of assessment for learning) has been that each page becomes a carefully crafted, beautifully manicured piece of artwork. The portfolio is still an art-object of its own, rather then being merely a mechanism through which the learner develops a prototype solution to the task they have taken on.

That is the difference with e-scape portfolios. They emerge dynamically through the sub-tasks that make up the overall activity. And the relentless time-clock ensures that the work for sub-task 1 (ST1) has just five minutes, then on to ST2 for another five minutes, and ST 3, 4, 5, and so on. The tasks range in time-scale from five minutes (minimum) to forty minutes (maximum) and there is just no time for second hand tidying-up. The entire focus of the project is on evolving a working prototype solution to the task. And as learners proceed through the task, the portfolio emerges automatically as the trail of work that is left behind through the development process. The portfolio is not an art-object but an evidence-trail. And the holistic judgements that are made of it through the 'comparative pairs' process accentuate the overview nature of the judging process.

RESEARCH

Judges frequently commented on the 'growth' of the ideas (from box 1 to box 20), illustrating the importance of the *evolutionary* state of work – rather than the presentation of it at any given point. This seems to me to be entirely healthy.

But – taking things a step further – imagine what would happen if learners themselves took a role in the judging process. Imagine looking at your own work – and your friend's work – alongside that of a complete stranger, and being asked to decide which was the more compelling piece of work. For classroom (learning) purposes, it doesn't matter a jot which way the decision goes, because what matters is not WHICH but WHY. What is it about this piece that is so compelling? Why would anyone think that *this* is better than *that*? Within this approach lies the potential for a real contribution to peer and self assessment for learning processes.

Teachers have always used exemplars of performance … mostly by keeping exemplars of excellent performance. They keep past portfolios precisely so that new learners can see and benefit from past years' successes. But in a new 'pairs' world, the exemplars are *immediate* and *current* and *multitudinous*. Thousands of pieces of work… all available at a click of the mouse… all with different approaches and nuances and attitudes and skills. It beggars belief to think that teachers and learners would not find ways of extracting valuable *learning* benefit from such a resource.

In conclusion

There are essentially two innovations in e-scape. First we developed a way of running six hour D&T assessment activities in workshops and studios – but in such a way that learners' portfolios emerged dynamically in a website. Having achieved this – and with 300 portfolios in the website – we developed the Thurstone/Pollitt pairs approach to assessment and showed that in this prototype form it could be done very reliably; far more reliably than current portfolio assessment systems can achieve. As a result of the political and policy interest that this has generated we have been asked to take the prototype to the next stage, and over the next couple of years TERU will be running phase 3 of project e-scape with DfES/QCA/Becta and Awarding Bodies. The challenge is to scale up the prototype to the point where it is capable of becoming a national assessment tool.

Within the broad ambit of this project (e-scape phase 3), the research questions that will steer it will range across pedagogic, technical, functional and manageability concerns. And one of the first specific explorations on my personal agenda is holism. We have shown time after time (from APU in the 1980s to e-scape last year) that design and technology teachers are very good at judging holistic capability. But being able to do it is not quite the same as being able to explain it. By the time we get e-scape phase 3 to the point of pairs-judging (Sept/Oct 2008) I would like to be able to explain holism and how it is that we can judge it so accurately.

If readers would like to know more about the work, or to get involved in the next phase, contact us at TERU (c.nast@gold.ac.uk).

References

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

Kimbell, R. (1997) *Assessing Technology,* Buckingham Open University Press.

Kimbell, R. Wheeler, T. Miller, S. and Pollitt A. (2007) *e-scape portfolio assessment – the final report of project e-scape (phase 2)*. Technology Education Research Unit. Goldsmiths University of London.

Pollitt, A. (2006) *Grading students' work,* an unpublished report to TERU for project e-scape. Published as part of Kimbell, R. Wheeler, T. Miller, S. and Pollitt A. (2007)

Pollitt, A. with assistance from Elliott, G. and Ahmed, A. (2004) 'Let's stop marking exams' Paper presented at the *IAEA Conference*, Philadelphia, June 2004.

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