The National and the Local: Conflicting requirements in the assessment of learners' performance

Prof Richard Kimbell, Goldsmiths, University of London

Introduction for the 2015 DATA Special Edition
This paper originates in a Keynote presentation at the Technology Education Research Conference (TERC2014) Sydney, Australia. Nov 2014. It arose as an invitation from the conference team to consider the tensions that arise from the very different concerns of formative and summative assessment. Specifically in this case the organizers were aware of the digital tools that we had developed for assessing learners' performance – and that we had shown these tools and approaches to be highly reliable. They were interested to hear how these tools and approaches might fit with the culture of learning in the classroom. Are they seen as an externally imposed discipline – or do they somehow contribute to an enhancement of the culture of the classroom?

Introduction
The formal assessment of performance in schools is typically undertaken by recognized authorities in assessment. In the UK this is sometimes private Awarding Organisations like AQA or Cambridge Assessment (eg for 16+ or school-leaving certification) and sometimes it is Government bodies like the Standards and Testing Agency (e.g. for National Curriculum Assessments). Arguably this latter is more about testing schools than children – using children merely as a lever to gain some purchase on schools' performance. In either event the priorities informing these assessments will be national. The tests must be deliverable and manageable nationally and the data must produce an articulated and reliable national standard. But the vast majority of assessments in schools are made by teachers, and they typically have other, more local, classroom concerns. Of course teachers are interested in how they measure up nationally, but they principally want to know how they can help their individual students to improve their performance. What are the strengths and weaknesses of the individuals and how might the teachers modify their classroom practices to enhance their learners' performance?

The assessment of performance is one of those fields where technical national requirements (for reliability and standards) meet local cultural practices (of pedagogy and individuality). And the meeting is frequently uncomfortable and unsatisfactory. In this paper I outline an approach to assessment that reconciles local cultural practices with national requirements.

Figure 1 Assessment framework for the 1970 Certificate of Secondary Education in Design Studies

The atomization of assessment
In the 40 years from 1970-2010 the process of assessment became ever more atomized. Whilst global judgements of quality were – at one time – seen as adequate and normal, it progressively became necessary to drill down into such judgements and provide more detail. As an illustration of this, in 1970, the assessment of a student design portfolio for the 16+ Certificate of Secondary Education 'Design Studies' was based on 5 judgements (each out of 5) summed to a single figure (out of 25). In 2010 a similar portfolio submitted for a National Curriculum assessment would be subjected to analysis against 150 'Statements of Attainment' that then have to be amalgamated through a complex set of rules to arrive at a final 'NC level'.

There were two reasons for this progressive atomisation. First was the belief that by identifying *elements* of, or aspects of performance, the final judgement might be more justifiably and more reliably decided. Second was the pedagogic priority, that identifying the *elements of performance* that are praise-worthy or inadequate makes it possible to see how the teacher can help the learner to improve his/her performance. The first we might call

atomization for reliability, and the second we might call *atomization for improvement*. (See Kimbell, 1997; Kimbell & Stables, 2007).

The articulation of assessment *criteria* became a fixation in the 1980 and 1990s; they proliferated into every facet of performance assessment. Along with the tendency went the belief that we were somehow transforming assessment from a personal view into an objective science. And this was despite the warnings of countless writers. Angoff, for example, observed that behind any criterion, there lurks a set of norms (Angoff, 1974), or Persig who argued that quality must be understood without definition; a direct experience independent of and prior to intellectual abstractions (Pirsig, 1991). Wiliam (1998) went so far as to suggest that most summative assessments were interpreted not with respect to criteria (which are ambiguous) nor with respect to norms (since precisely-defined norm groups rarely exist), but rather by reference to a shared *construct of quality* that exists in well defined communities of practice.

But against the in-rolling tide of criterion-creators, these were voices in the wilderness and the tide continued to roll in. However generously we might wish to judge the motives of those responsible for this trend towards atomisation, the *effects* of it – the outcome (at least in England and Wales) – has been utterly disastrous. And the scale of the disaster can be judged by reference to two events. In 1992/3 the burden on teachers (person hours and paperwork) of National Curriculum assessments had reached such a level that there was a completely unprecedented national boycott of all assessment by teachers. It was so absolute and so widely supported (including by heads and parents) that in the end the Minister responsible was sacked. Some modest fiddling at the edges followed, but no change of any real significance resulted. So a broken system was patched up and hobbled on. Then in 2006 the new Minister (we had 19 between 1970-2010) decided that the assessment of student portfolios of coursework (e.g. for 16+ GCSE design & technology) was so unreliable that the whole process had to be abolished. Coursework assessment was no longer an acceptable mode of examination.

I should make it clear that there were plenty of other disastrous events accompanying the atomizing trend, but I have chosen to identify these two because they provide an interesting juxtaposition with the motives underpinning the trend. Atomization for the purposes of classroom improvement generated an entirely opposite effect; an absolutely solid boycott from classroom teachers. They wanted nothing to do with it, claiming that (i) it was massively burdensome and (ii) the assessment told them nothing they didn't already know. Atomisation for the purposes of improved reliability resulted in another entirely opposite effect. It generated such chaotic unreliability that the Minister banned coursework assessment. Forty years of progressively atomized assessment created uncountable hours of hard labour for teachers and hopelessly unreliable outcome statistics. By any standards, the end result of this atomizing trend was catastrophic.

When you find yourself in a hole…stop digging. It is surely time to change direction and explore new and less flawed models of assessment.

Re-thinking assessment

In 2004 we had been awarded a new research project (e-scape) in TERU at Goldsmiths – to explore the possibility of on-line portfolios and digital assessment. It proved to be a 6 year venture through three phases amounting to an investment of approximately £1m. By the time we got the contract, the writing was already clearly on the wall about the existing discredited model of assessment (long lists of criteria, all scored and added-up by the students' own teachers) – and we understood that this new project gave us a license completely to re-think what assessment might be like in a digital world.

Within the project 15–year-old learners constructed digital portfolios of work (in design & technology, science and geography) in response to authentic, extended tasks. These tasks were conducted in normal design studios, science labs, and on geography fieldwork. In design & technology, learners designed and developed products using PDAs as digital sketchbooks, notebooks, cameras, and voice recorders (NB this was in 2005/6, well before ipads and other tablet devices became available). Their work was automatically and simultaneously sent through a wi-fi connection to a secure web-space in which their virtual web-portfolio emerged. At the end of the national trials, we had 350 design & technology portfolios, 60 in science and 60 in geography. (See Kimbell et al. 2009, Kimbell & Stables 2007)

From the outset of the project we realized that the web-based nature of the portfolios enabled us to explore a radically different approach to assessment. And the approach was informed by three big ideas.

Three big assessment ideas

1. Absolute or comparative judgement

School-based assessments typically use numbers on a scale. Judge the portfolio against this criterion on a scale

of 1-20 or 1-8 (depending on its perceived importance). Assessment is on an absolute scale, and – theoretically – if I award 7/20 to student x, then that work is exactly the same standard as student y in another school where another teacher has also awarded 7/20.

But, judging on absolute scales is VERY difficult. How warm is your current room (in degrees C)? How heavy is the book you are reading (in grams)? How fast are you driving (in mph)? We typically do not hold a standard against which to measure these judgements – so unsurprisingly we are more often wrong than right.

> When someone comes to make a judgement in the everyday world, 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 this difference 'a point of view'…. All judgments are comparisons of one thing with another …the judgment depends on what comparator is available.
>
> (Laming, 2004, p.17)

> When we try to judge a performance against grade descriptors we are imagining or remembering other performances and comparing the new performance to them. But these imagined performances are unlikely to be truly representative of performances of that standard, and very likely to vary in the minds of different judges.
>
> (Pollitt, 2004, p.7)

What we can do…. VERY reliably…is to weigh one thing against another. As Laming says, 'All judgments are comparisons of one thing with another' and if I am asked to compare the temperature of two adjacent rooms I can immediately tell you which is warmer even though I can't tell you the 'real' temperature. Or two books…which is heavier. Or two portfolios…which is stronger. Comparative judgement is easy and accurate.

2. Judging parts or wholes

I have already spent a while deploring the trend towards atomization in assessment, but it is worth thinking for a minute about the *reverse* of atomization. Imagine that you are a biologist presented with a new species of plant/animal that you have been asked to identify and classify. And you have been provided with a set of instruments including a microscope, a hand lens and a ruler. What would be your procedure?

I'm not a biologist, but simple common sense suggests that you start with the naked eye 'its 25 mm long with a body in two parts and 6 legs'. You might then pick up the hand lens to get a better look at what appear to be the eyes. Then you might need a microscope to see how the scaly surface of the body is composed. What I would definitely *not* do is to ask for 150 microscope slides of bits of the specimen and – on that basis – try to identify what it is.

We start from big pictures and - progressively - drill down for more detail. We do not start with a box full of details and try to build up a big picture. So why on earth do we do that when we are trying to assess a student's performance? Our first instinct should be to say 'this is a great piece of work' or 'this is really weak' and then - progressively – drill down into it to find out why.

Holistic judgement has long been understood to be important in design & technology. Indeed, in the 1988 Interim Report of the D&T National Curriculum Working Group (a year before the full report was published) they commented as follows:

> 'These considerations point to the conclusion that, because Design and Technology activity is so integrative the approach to the assessment of pupils' performance in this area should ideally be holistic'
>
> (DES, 1988, para 1.30)

It was a matter of some astonishment therefore when the full National Curriculum report was published a year later and proposed an assessment regime involving ticking boxes (or not) against 150 atomised Statements of Attainment. All the available evidence advised the reverse approach.

When we were running the APU Assessment project at Goldsmiths for the Department of Education and Science (1985-1991) we had about 120 teachers involved in the assessment of the student work that was generated in the 1988 testing programme. We asked these teachers to make an initial *holistic* judgement (on a 6 point scale) and then follow it with a series of increasingly detailed judgements of elements of the work.

> Of all the judgements markers made, they felt more confident and were more reliable when assessing holism.
>
> (Kimbell et. al., 1991, p133)

3. Sorting networks

A sorting network is a mathematical approach to sorting a sequence of numbers. Sorted data (e.g. in a computer where files are sorted by file-size or date) is much easier

The National and the Local: Conflicting requirements in the assessment of learners' performance

Figure 2 Computer Science Unplugged YouTube 2005 (https://www.youtube.com/watch?v=30WcPnvfiKE)

to work with than unsorted data, so mathematicians have spent a long time working out protocols for sorting. It's difficult for me to explain the working of the sorting network that I'm interested in for this paper, so the best way forward is to watch a short you tube video (only 2 minutes) that outlines the approach. It shows 6 children sorting a set of numbers into order by following lines on the floor and (when they meet another) going left if their number is smaller, and right if its bigger. Magically, the numbers sort themselves into order.

Assessment is a sorting problem. We start (as a teacher) with a pile of essays and end up with a sorted pile (best to worst). Awarding Organisations start with a random mass of candidates and end up with a sorted candidate list. Once the work to be sorted can be readily distributed and accessed (which they can when its web-based) then enough people can get involved to undertake the sorting process.

Assessment in project e-scape

Having established project activities in design & technology, science and geography, and having derived the web-portfolios, we turned out attention to the problem of assessing them. And after a series of experiments we embarked upon a completely new approach to assessment that used the three big ideas outlined above (comparative judgement, holistic judgement and sorting networks). Wiliam would describe the approach as 'construct-referenced' assessment (Wiliam, 1994) in which performance is not defined in advance as a set of learning outcomes, but rather the *construct of quality* that underpins assessment judgements is sufficiently understood and shared by a community of practice.

In practice, comparative judgement requires that scripts (portfolios) are sent to judges in pairs, and the judges simply report which one is the 'better' in each pair. They make this judgement informed (in our case) by five headline criteria. But they don't judge the criteria separately. They are asked to hold these criteria in mind as they make their holistic judgement. Whilst current 'marking' approaches require only that each portfolio be scored once, comparative judgement needs each portfolio to be seen several times in different pairings.

> 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 internalized 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. The result of comparisons of this kind is objective relative measurement.
>
> (Pollitt, 2004, p.6)

In addition to the reliability benefit of the canceling out of judges' individual bias – a related benefit was immediately clear. Conventional marking is by one marker of one portfolio (at a time). The whole process is individualized. With comparative judgement – using web technologies – it is possible to have whole teams of judges sharing their judgements about the whole sample of portfolios: a collective process that also contributes to the improvement of inter-rater reliability.

In the final phase of the e-scape project we automated this paired judgement process by developing the 'adaptive comparative judgement' (ACJ) engine, a Rasch modelling algorithm that identified the portfolio-pairs to be judged next, and which judge they should be sent to. It is an *adaptive* algorithm; it learns about the portfolios as it accumulates judges' responses. So at the outset a judge might be sent two portfolios that are randomly chosen from the sample, and if one was pretty good and one fairly weak it's an easy judgement to decide which is stronger. But gradually the engine works out an approximate rank-order for the portfolios, so it can send judges a pair of portfolios that are much closer together in the rank. This refines the rank very rapidly.

In the 2009 national trial – with 350 portfolios and 28 judges – we rapidly arrived at a rank order with a reliability statistic of 0.95. This is an astonishing statistic. Absolute reliability about a set of multi-media portfolios that portray creative designing activity by 350 learners. Never before

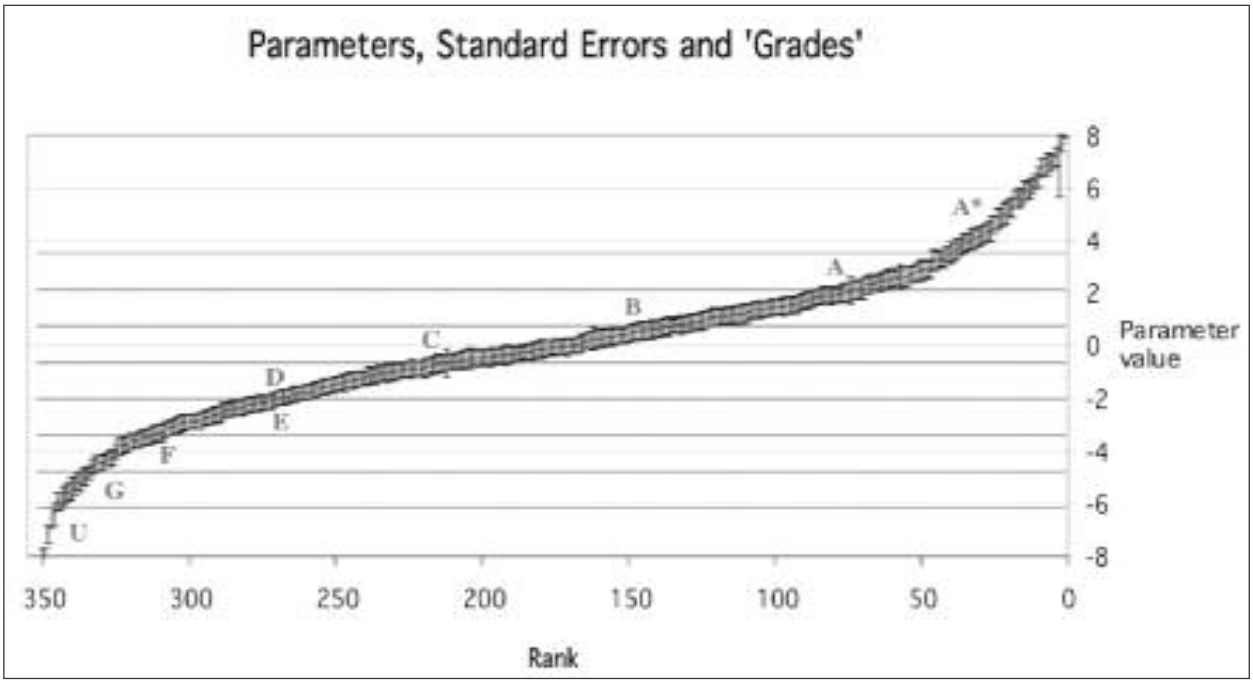

Figure 3 The rank created by Adaptive Comparative Judgement in the 2009 national trial.

has it been possible to produce this level of reliability with such data. And all the conventional paraphernalia of assessment was gone. No extended scoring sheets…no allocation of marks and painful calculation ofoverall scores…no 2nd markers and disagreements…no moderation.

All the judges had to do – in relation to each of the pairs of portfolios sent to them – was to say 'this one is better than that one'. End of story. Our teacher/judges thought it was wonderful and were delighted that they had contributed to such an astonishingly reliable outcome.

Why is this model of assessment (ACJ) so reliable?
There are at least five reasons why the assessment of performance (using comparative judgement, holistic judgement, and sorting networks) is so much more reliable than conventional marking.

1. Because it is collaborative judgement
In a normal marking context, teachers are responsible for marking the portfolios of their students. So teacher x in his school marks all the x portfolios and teacher y does the same in her school. What is the shared understanding of teachers x and y? Do they hold a common standard? Sometimes they will and sometimes they won't. And that is just not a good enough basis for deciding which students should pass and which should not.

With the comparative judgement process using ACJ, all the teachers submitting students for the examination become judges. All the portfolios are held in a big national pot (actually in a server-farm under Canary Wharf in London). So all the portfolios are mixed up together and not held at the school level. Judges are sent a pair to decide upon – and then another pair – and then another and so on. In the 2009 trial, each of the teacher/judges made approx 120 paired judgements and that was the end of the process. On average, judges took 4 mins to make a judgement, so 8 hrs in total. And most reported that it was a shorter time than they would normally have spent on marking their class of portfolios in the normal way.

Critically however the teachers were not judging their own class work – they were contributing their judgements to the whole national pot. So those teachers – for the first time – could see what the national standard of work was really like. We asked for feedback on the process….

> The judging system feels to be fair; it doesn't rely on only one person assessing a single piece of work. It removes virtually all risk of bias.... It feels safe knowing that even if you make a mistake in one judgement it won't significantly make a difference to the outcome or grade awarded to the student as other judges will also assess the same project. Also knowing that the system automatically checks the consistency of the assessor's

judgements again reinforces the feeling of fairness that this process brings. (DW) much, much faster…less scary (re individual marker impact on individual learner life chances)…get a whole view much more readily (RW)

(Kimbell et. al., 2009, pp.69-72)

Figure 4. Teachers marking a class set of portfolios

2. Because it is comparative not absolute judgement
For the first time in a national assessment process, teachers were not being asked to stick a number against a set of criteria. They just had to look through both portfolios – consider the basket of criteria we have trained them to identify – and then make a single holistic judgement. Is it portfolio x or portfolio y? The overriding reaction of the teachers was astonishment at how easy it all was

Easier assessment; no need to calculate grades and points (RM)

Speed of judging (VG)

It's worth pausing for a moment to consider the contrast with their normal process of assessment, for we were surprised how readily they took to the idea of comparative judgement. When we discussed it with them it all became clearer. Normally they start by laying out their portfolios (best-worst) on the desks in a room and then they go round the room (often in teams) filling in the forms to get the final marks. And this becomes a comparative judgement process. "We've given Julie 7 for that, and John is definitely weaker" They are using the benchmark they set for Julie as a means for deciding on the mark for John. It might look like criterion-based judgement, but it's also comparative.

And the big difference with ACJ is that what emerges is not a mass of different school-based standards, but a single national standard to which every teacher has contributed.

3. Because it is holistic not atomized judgement
The teachers where absolutely unanimous about the importance of holistic judgement – and its clear advantage over the atomized approaches with which they have become so wearily familiar.

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. (AM)

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. (DP)

One of the major strengths of holistic judgements I see is its flexibility…in which you can give credit to students for what they have actually done rather than whether they are able to "tick the boxes" to match a set of assessment criteria. (DW)

Making holistic judgements meant that I was not forced to give credit to an apparently well-designed project that was completely unrealistic in terms of being an actual product. (VG)

(Kimbell et al 2009 pp 69-72)

4. Because the algorithm underlying ACJ is very efficient
Given 350 portfolios and the principle of comparative judgement, one might think that every portfolio has to be compared with every other one. That is 350 x 350 judgements! In reality the algorithm does a lot of the work for us and it works on the idea that if A beats B which beats C which beats D…then A will probably also beat D. And it works out a probability for that. Imagine a matrix of 350 x 350. The boxes in it are where judgements are made (yes or no – based on which wins). And the trick with the algorithm is how many of the boxes can remain empty (those two portfolios have not been directly compared) and yet produce a reliable outcome.

In the initial ACJ prototype for the 2009 trials we worked on the notion that each portfolio would need to be compared with 20 others. But in the event – after only 16

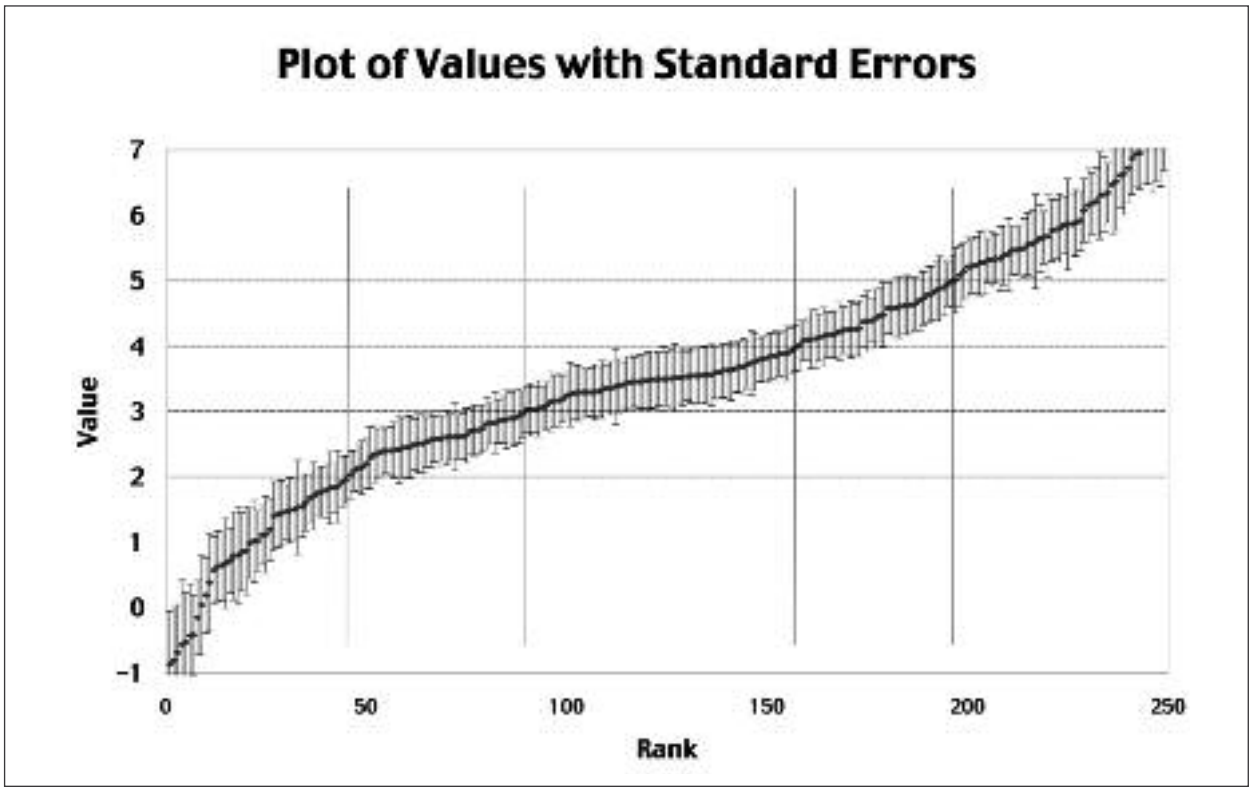

Figure 5. The standard error 'tails' show us which portfolios are causing disagreement

rounds – the rank order didn't change however much more judging we did. Then, with some further refinement, the algorithm produced a solidly reliable rank after 11 rounds. And currently (Oct 2014) it requires 10. And remember that the reliability is far higher than anything that can be generated using existing marking processes.

> ...the portfolios were measured with an uncertainty that is very small compared to the scale as a whole…. The value obtained was 0.95, which is very high in GCSE terms. Values of 0.9 or so are considered very strong evidence of validity for the test.
>
> It is worth noting that the average standard error of measurement for a portfolio was 0.668, which is just less than half the width of one of these "GCSE" grades. It is unlikely that many GCSE components – or any that are judged rather than scored quite objectively – could match this level of measurement accuracy.
>
> (Pollitt, in Kimbell et. al., 2009, p.81)

5. Because the problems encountered are made explicit
When teachers make judgements about portfolios there are two potential sources of problems for the judgement. First, the teachers may make random or inconsistent judgements and second, any given portfolio might (for some non-obvious reason) cause judges to disagree. In both cases the ACJ engine collects the data to decide what we might do about it.

All the judges generate a 'misfit' statistic that tracks the consistency of their judgements against those of the whole judging team. It might be thought of as a 'consensuality' measure. If a judge is making judgements that are way out from those of the rest of the team, we need to know – and to understand why. They may be right and the rest of us wrong – but we'll never know if we don't check. So the misfit stats accumulate for every judge and during the 2009 trial – with 28 judges – only three approached anywhere near to a misfit score that required intervention.

As for the portfolios – they too accumulate a misfit score that shows as the 'standard error' attaching to each portfolio. The blue dots that make up the blue line are the 'real' position of the portfolios, but the grey 'tails' either side indicate the size of the standard error on each. Some have bigger tails than others and if they become too big they can be pulled out and subjected to a separate moderation process.

So if the teachers and the portfolios might be sources of error – both are covered by the internal processes of the ACJ engine.

Conclusion

I suggested in the introduction to this piece that the assessment of performance is one of those fields where technical national requirements (for reliability and standards) meet local cultural practices (of pedagogy and individuality). And I suggested that the meeting is frequently uncomfortable and unsatisfactory. One has only to see the increasing number of appeals by schools against decisions by Awarding Organisations (at least, in England) to gauge the extent of the misfit between the concerns of teachers and those of national assessment agencies.

> **Sharp rise in appeals against primary school exam results**
>
> Rising numbers of primary schools lodged official complaints over marking in SATs tests this year amid fears children may have been given the wrong grade, it emerged today. Some 5,537 reading and maths papers were sent for review in the summer – an 88 per cent increase in just 12 months. The process resulted in 1,255 exam scripts being marked up.
>
> (The Telegraph 31st Oct 2013)

I promised at the start of this paper that I would outline an approach to assessment that reconciles local practices with national requirements, so it is time to make good on that promise.

In 2010, the national assessment Standard Assessment Tasks (SATs) (in England) were managed by The Qualifications and Curriculum Authority (QCA), who were already alarmed at the rising number of appeals against the judgements made in their SATs for 11 yr olds. They were particularly alarmed in the 'writing' tests where appeals had sky-rocketed. Because the QCA had been responsible for monitoring the progress of project e-scape, they knew of our work with ACJ. Indeed only the previous year (2009) we had submitted to them our phase 3 report – detailing the process and the result. So they asked whether it would be possible for us to use the ACJ methodology for the assessment of pupils' writing. When we said that we could, they provided a sample of 1,000 scripts. Each was of a piece of free story-writing (between one and two sides of A4) on a given theme. We adapted the ACJ interface to take the written text and recruited 54 primary teachers to do the judging. The result was as successful as we had expected and the teachers' response was also as predicted.

> The overall reliability of the assessment was 0.961, meaning that this assessment was considerably more reliable than any other assessment of writing that we can find reported in the national or international literature.
>
> When the judges were asked for their opinions about the method, they listed these main advantages: speed, the holistic nature of the process, increased fairness, professionalism, and a positive impact on teachers and schools.
>
> Every respondent described it as Fine, Easy, or Very easy.
>
> When asked if they would prefer to use the Comparative Judgement method or return to Marking, 25 chose Judgement, 0 chose Marking, and 2 voted for both.
>
> (Pollitt, Derrick and Lynch, 2010, Summary)

Moreover, in the section of the report where we invited teachers to feed back their comments, we received observations that were almost identical to those resulting from the 2009 e-scape trials.

> Each script being judged by many professionals instead of a child's fate resting on one marker
> Fairer with many assessors
> Reduces subjectivity in marking as it isn't based on just one person's opinion
> It takes the pressure off being the sole person responsible
> Allows scripts to be considered in their entirety without individual features assuming priority because of a mark scheme
> Judge the whole piece
> It feels natural and fair
>
> (Pollitt, Derrick and Lynch, 2010, Sect 3.1)

But I would particularly draw readers' attention to the teacher comments that centred on the professionalism of teachers and the extent to which the approach would make a positive impact on teachers and schools. There were many comments of this kind.

> Allows for professional judgement
> Uses our years/ decades of experience
> This system makes more sense – making a general judgement as to the level of a piece of work is what most teachers do anyway before they go through the criteria to prove what they think
> As a teacher, I felt I was able to make a better judgement in terms of the child's overall approach to texts and it excites me to think we could actually teach

> children the overall value of texts rather than subject them to judged deconstruction of a text.
> (Pollit, Derrick and Lynch 2010, Sect 3.1)

These comments – about professionalism, normal classroom practice, and exciting teaching opportunities – are not the kinds of responses one expects to hear from teachers just emerging from an extended bout of marking. But this was not marking. This was Wiliam's (1994) community of practice articulating their shared construct of quality.

So reconciliation is possible between good, professional, teacher expertise (the culture of the classroom) with the needs of national assessment (reliability and standards). The one does not exclude the other. And moreover I would leave readers with a final observation.

The Awarding Organisation that initially marked the writing SATs had the normal extended hierarchy of subject officers, examiners, chief examiners, moderators and senior moderators. And still they managed to produce such a suspect result that thousands of schools appealed the outcome. In our ACJ trial of the very same writing SATs, not only did all the teachers collaborate in arriving at a *common* standard, but moreover the process was judged to be professionally worthwhile for them. *And there was no hierarchy of examiners*. We had one expert analyzing the misfit statistics and checking the reliability as it emerged – but the entirety of the judging itself was in the hands of the community of practice; the classroom teachers. As it should be, since they are the people who taught the children to write their stories in the first place.

Do not underestimate the significance of this. If this democratised model of construct assessment were to be adopted nationally and internationally, it would dramatically empower classroom teachers – enabling them to develop and share their constructs of quality in learners' work. And at the same time it would equally dramatically improve the reliability of national assessments.

References

Angoff, W. H. (1974). *Criterion-referencing, norm-referencing and the SAT. College Board Review*, 92 (Summer), 2-5, 21.

DES/WO. (1988). *National Curriculum Design and Technology Working Group Interim Report*. London: Department of Education and Science and the Welsh Office.

Kimbell, R., Stables, K., Wheeler, T., Wozniak, A., & Kelly, A. V. (1991). *The assessment of performance in design and technology*. London: SEAC / HMSO.

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

Kimbell R & Stables K (2007) *Research Design Learning*, Dordrecht, Netherlands, Springer.

Kimbell, R., Wheeler, T., Stables K, Shepard T Martin M Davies D. Pollitt A Whitehouse G (2009). *e-scape portfolio assessment Phase 3 Report*. to SEAC. London: Technology Education Research Unit, Goldsmiths University of London.

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

Pirsig, R. M. (1991). *Lila: an inquiry into morals*. New York, NY: Bantam.

Pollitt, A. (2004, September). *Let's stop marking exams*. Paper presented at the IAEA Conference, Philadelphia.

Pollitt A, Derrick K, Lynch D, (2010) *Single Level Tests of KS2 Writing: the method of paired Comparative Judgement*. TAG Developments. Unit 3:11, Canterbury Ct. Brixton Rd London SW96DE

The Telegraph on-line (31st Oct 2013) *Sharp rise in appeals against primary school exam results* http://www.telegraph.co.uk/

Wiliam D (1998) *Enculturating learners into communities of practice: raising achievement through classroom assessment*. Paper presented at European Conference on Educational Research, Ljubljana, Slovenia, September 1998.

Wiliam, D. (1994). *Assessing authentic tasks: alternatives to mark-schemes*. Nordic Studies in Mathematics Education, 2(1), 48-68.