IAEA 2009 Assessment for a creative world (13th-18th Sept 2009 Brisbane)

project e-scape: creative performance … assessed with creative spectacles

Professor Richard Kimbell Goldsmiths University of London <r.kimbell@gold.ac.uk>

Abstract

Creative performance lies at the heart of every school subject. Innovative investigating in science; imaginative spatial exploration in geography; creative product development in design. Attempting to capture these creative performances for assessment purposes has frequently destroyed them. Just as the illusive butterfly is damaged or crushed by crude wafting of a catch net.

Designing activities that enable us to reveal and capture creative performance has led us to explore a mass of digital technologies – and to the evolution of 'real-time' portfolios that illuminate collaborative as well as innovative performance. The activities move through a series of sub-tasks enabling learners to leave behind them an evidence-trail of their route through the task. The performance is all captured in real-time web-portfolios.

At the assessment end of the process, creative performance has typically been ripped apart by the application of atomistic criteria. Yet teachers know which of their students are the really imaginative scientists; the innovative designers; the eloquent authors. When the bits don't add up to the right answer – confident teachers change their 'bit' scores to make sure it does.

Holistic judgement circumvents this time-wasting nonsense. Criteria are used to inform the judgement – but the judgement itself is of the whole integrated performance. To be taken seriously in high stakes assessment, such an approach has to be especially alert to technical challenges. With Pollitt we have created a 'pairs engine' that automates a comparative pairs (Thurstone) judging process. Not only is the reliability of the assessment extraordinarily high (0.95) but the data on individual portfolios and judges allows us to identify and deal with any technical problems.

Overview of project e-scape (2004-2009)

The e-scape project has progressed through three phases and is built on a series of innovations in the world of performance portfolios and their assessment. Phase 1 (2004-5) was a 'proof the concept' in which we explored a range of hand-held digital tools and systems to establish that it was possible to link them directly to web-portfolios to capture live classroom 'performance'. In phase 2 (2005-7) we built a prototype system for portfolio assessment in design & technology based on learners' performance in a 6 hr design task. The prototype worked very effectively and in phase 3 (2007-9) we have established the transferability of the system for assessments in a range of subjects and the scalability of the system for coping with and managing national assessments.

There are two principal innovations in the e-scape system.

First, we have created a system in which school learners use hand-held digital tools in the classroom to create real-time web-portfolios. The hand-held tools are linked dynamically to their teachers' laptop – operating as a local server. This sends a series of tasks to the learners and 'hoovers-up' anything that they produce in response to them. Learners' response can be in text / draw / photo / audio / video / mindmap / SS. The local server is enabled to upload – dynamically (in real time) - all the data from a class/group into a website where learners web-portfolios emerge. The files are automatically converted into Flash for effective file management in the portfolio.

Second, we have created a web-based assessment system based on a 'Thurstone pairs' model of comparative assessment. The web-based portfolios can readily be distributed anywhere at anytime – enabling multiple judges to scrutinise the portfolios simultaneously. The judging in phase 3 involved 28 d&t judges, 6 science and 6 geography judges. All the judging was completed on-line in a short time-window and with extraordinarily high reliability (0.95). All these data can be downloaded into existing Awarding Body back-end systems for awarding and certification purposes.

Elements of the e-scape system

i) the authoring tool

To enable the e-scape system to be as flexible as possible, we have developed an activity-authoring tool to enable teachers, examination bodies, or researchers to design their own activities. The tool is available on-line. Essentially the tool allows teachers to decide on any number of possibilities:

- the content area of the task (eg science/geography/design/drama)
- the overall duration of the activity
- the time sequence (eg 6 one-hour sessions, or a single 3 hr block)
- the sequence of sub-tasks that build up the overall portfolio
- the response-mode of learners (eg drawing/writing/photo/audio/video)
- the degree of flexibility in the timing of sub-tasks (controlled>flexible)
- the resource materials to be embedded (eg texts / images)

The tool allows teachers to design the activity – and modify it in the light of 2nd thoughts or trial runs. It enables different sub-tasks to be selected for different learners – allowing teachers to personalise the activities to particular learners and their needs. Equally however for assessment activities, examination bodies can ensure that exactly the same activity is presented to all learners in all test schools.

ii) the exam management system (EMS)

Once the activity has been designed, it can be transferred into the EMS, and it is from here that teachers manage the activity in the classroom. The EMS runs in the teacher/administrator's laptop which operates through a local area network with wi-fi connectivity to learners' hand-held devices in the classroom / studio / workshop / laboratory.

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

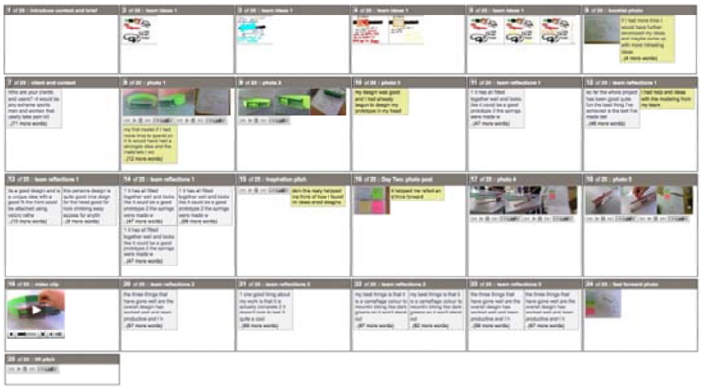
At the start of the activity, the teacher activates sub-task 1 and this is sent to learners' devices. They work on it for the designated period (writing, drawing, taking photos etc), at which point the teachers laptop 'hoovers-up' all their work back into the EMS and sends the 2nd sub-task to learners. The closed network guarantees good data transmission between the teachers and learners devices. Throughout the activity, the EMS enables the teachers to check that all the learners' devices are connected and operating properly – and gives a simple visual check on the battery state of each device.

The teacher works through the task merely by clicking on the big green 'next' arrow that takes the activity forward to the next subtask. Alternatively the teacher can choose to by-pass some of the sub-task steps using a drop-down menu that lists all the sub-tasks that were designed into the activity. At the end of the activity, the teacher uploads all the data from the EMS into the portfolio management system.

In the June 2008 e-scape round, teachers in 19 schools operated these activities with classes of 21 learners. The activities were in science (3 hr [one morning] activity), design & technology (6 hr [two mornings] activity), and geography (5 hr [whole-day] activity).

iii) portfolio display

From the activities in the 19 schools, 350 d&t portfolios, 60 science and 60 geography portfolios were uploaded into the web-space. The portfolios reflect the sub-task structure of the activities, but the display of them can be customised.



For the purposes of the design & technology phase 3 portfolios, we intended to use a pairs judging process with multiple judges. We optimised the display with 'thumbnail' images so that all 25 sub-task boxes would fit as a single screen (without scrolling) on 20inch monitors. This portfolio-at-a-glance arrangement allowed judges to scan the whole piece of work. Additionally however, judges can click on any of the thumbnails (eg photos / drawings / sound files) that automatically jump to full screen images or play as voice/video files.

iv) the pairs engine

The pairs engine manages 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) and is often used for inter-board reliability studies for GCSE and other school-based examinations. Pairs judging in these cases is used to check the reliability of the assessments that have already been made. In 2008 for e-scape phase 3 we have developed the system further so that the pairs judgements are the ***front-line assessments*** of the portfolios (ie there is no other assessment). We have developed the pairs engine to run this as an automated process.

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 & 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 6. Whilst training sessions for judges were conducted face-to-face in free-standing training days, the judging subsequently took place remotely – typically in judges’ homes. We had judges logged in from Ireland, Israel, and from across the UK.

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 in part explained by the fact that each portfolio is compared to approx 20 others and is seen by many judges. The same levels of reliability were achieved with the science and the geography judging teams looking at their portfolios.

The out-turn data from the pairs engine can be fed into Awarding Body back-end systems for any subsequent awarding processes.

Issues arising

a) the authentic voice

One of the most compelling elements of the e-scape experience has been the response of learners to the activities and their reactions to the portfolios that they generate. The activities are choreographed through sub-tasks involving many kinds of response (text /photo etc) and the authentic voice of the learner emerges in the resulting portfolio, which is a rich portrayal of the experience of the activity.

The portfolios often have powerful evidence or meta-cognitive processes at work – and particularly in the voice files. We have embedded these at points through the activity – and typically we ask learners to reflect (for just 30 secs or a minute) on their work so far: how they think its going and what they plan to do next. It is astonishing what learners say … things that they would never write down. And in the process they provide far more than just descriptive data of *what* they have done. They discuss their rationale for *why* they have done this or that and they even express their feelings and *attitudes* about the task and their work. These rich layers of data enable the portfolios to go way beyond conventional probing of learner capability. Scanning through them, it almost feels as if one is sitting alongside them in an informal viva setting.

Teachers are enthusiastic about the ability to capture creative performance.

> What a powerful tool for a teacher! I particularly like the way sound files, photos and videos can be simply integrated into a normal classroom situation, making it possible to access thoughts and ideas of every individual (including SEN) rather than just those who it is possible to get round and talk to - really powerful for assessment for learning. Tremendous potential for all curriculum areas and age groups, including cross-curricular work.
> teacher 11 (Kimbell et al 2009 p105)

b) creative performance

The concern that initiated the e-scape project was that current assessment arrangements in the UK have become so formulaic that they have rendered the notion of creative performance in an examination almost meaningless. Learner portfolios – in geography, science, history, music or design – are typically **not** portrayals of the creative process at work. Rather, they are the 2nd/3rd/4th hand re-constructed efforts (by both teachers and learners) to hit all the buttons in the assessment rubric. And the linked tragedy is that, since too many really creative youngsters are simply not prepared to play that game, their performance gets marked down. The top marks go the learners with the dedicated patience to produce a neat 5th version of their 'creative journey' in the portfolio, and who have teachers who know the rubric by heart.

We do well to recall that the rhetoric of Ministers has long been urging on us the importance of creative performance.

> "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 T 1999)

But the demands of what Taylor and Hallgarten (Institute for Public Policy Research) call the 'regulatory state' are such that deeply conservative traditions of regulation are seen to be required.

> It is difficult to see how the process can avoid dampening the growth of a culture of innovation and experimentation. The expanding regulatory State appears to be more about central enforcement than front-line empowerment.
> (Tayler M & Hallgarten J 2000 p10)

For school examinations, this plays out into equally conservative Awarding Body practices, and (in the context of assessed learner portfolios) defensive practices by teachers doing their best to ensure that their students get every mark that is possible. All of this is understandable, but the outcome runs in entirely the opposite direction to that expressed in the political rhetoric.

E-scape was conceived as part of a solution to this problem. Part of the solution lies in the ‘real-time’ nature of the portfolios. In the science task, activity progressed through a series of 15 sub-tasks over 3 hrs. At the outset learners are presented with the idea of road safety through video-snippets of traffic accidents from the Department of Transport. They are asked initially (for just 5 mins) to speculate and jot down some questions that they might ask to find out more about why they had happened. Then – in a subsequent 5 min session – they are asked to look through their questions and see which of them might be *scientific* questions, and to see if they can work on them to make some really good scientific questions. As each sequence passes, the data is locked in and embedded in the website. What emerges is the real unfolding story of the learners’ progress through the task. There is no going back … except through regular reflection sessions in which learners tell us what they are doing, why, and what they might do next.

c) holistic judgements of capability

All teachers in the e-scape trials volunteered to be e-scape judges. We ended up with 28 in design, and 6 in science and geography where there were many fewer portfolios.

The process of judgement is not of course criterion-free, since we have to have ways of describing what we mean by poor / better / best performance. The training day for judges started by enabling them to scrutinise exemplar portfolios in relation to (in design) 4 key criteria. They were asked not to score them, but rather to hold them in their head as a *collective* description of capability. They then debated and practised the exercise of holistic judgement. Their task (in the pairs engine interface) was to determine whether portfolio A or portfolio B represented a better overall (more capable) performance.

It did not take long for judges to become familiar with the engine interface, since it operates at a fairly intuitive level. However the ***digital portfolios*** offer such rich data sources that it takes a while to learn effective ways of navigating around them.
A month later, at the end of the judging process (having completed their 127 paired judgements) we asked all judges to complete a questionnaire about the whole process. The strengths of the e-scape e-portfolios were characterised in three ways. *First* in terms of presenting the ‘big-picture’ of learner performance; *second* in terms of capturing the critical ephemeral evidence alongside the more obvious artefact evidence; and *third* in the ability to see the growth of ideas through the activity.

> Portfolios displayed in this way have a huge advantage in that “the big picture” can be seen immediately. It’s very easy to get a “feel for the project” that would not be possible unless

> the whole project was displayed on one single screen. The ability to "dip in and out" of the different sections enabled me to reinforce my holistic mental picture of the project. (DW)
> The ability to hoover up ephemeral evidence of designing and creativity (TL)
> Improvement of ideas (VG)
> I could see progress / or when students had simply abandoned one idea and started another. It was also evident whether ideas had 'grown' and what inspired the students. (HW)
> (in Kimbell et al 2009 p 109)

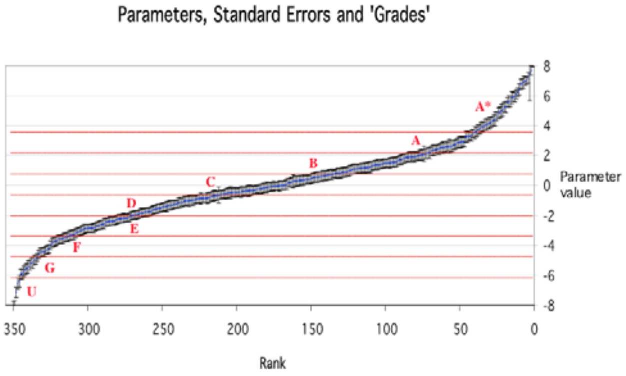
We were very aware that a critical difference between e-scape judging and GCSE marking is the centrality of the **holistic judgement**. We were pleased that judges felt they could hold a sense of holistic capability and use it to reward good performance.

> 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) of the parts of the process. (DP)
> (in Kimbell et al 2009 p 110)

d) assessment reliability

For the last four years of the e-scape project we have been working in association with Alastair Pollitt who first introduced us to the notion of Thurstone Pairs judgement. Our pairs engine (Lynch, Wheeler, Pollitt, Kimbell, Derrick: Patent No GB 0909539.9) is one of the outcomes of this collaboration.

In relation to the 2008 samples (350 in design & technology and 60 each in science and geography) Pollitt's analysis of the out-turn of the judging is both detailed and revealing



Concerning the design of the algorithm in the pairs engine:

> In effect, the Swiss tournament system [the 1st 6 rounds of judging] achieves an approximate targeting at almost no computational cost.

Concerning quality control:

> Because every single judgement made can be compared to the outcome predicted (with the benefit of hindsight) from the final rank ordering, very detailed monitoring is possible of the consistency of the judgements made by each judge, and of each portfolio.

Concerning the 'fit' statistic for judges

> Theory predicts that this statistic should average 1.00, and in these data it does exactly that the calculation gives 1.64 as a criterion [for fit], and only one judge [out of 28] exceeds this. It may be significant that this judge made only 59 judgements, while the others averaged almost twice as many. [NB in fact the judge was reluctantly forced to withdraw from the process for personal reasons]. Overall the amount of misfit seems quite acceptable.

Concerning the ‘fit’ statistic for portfolios

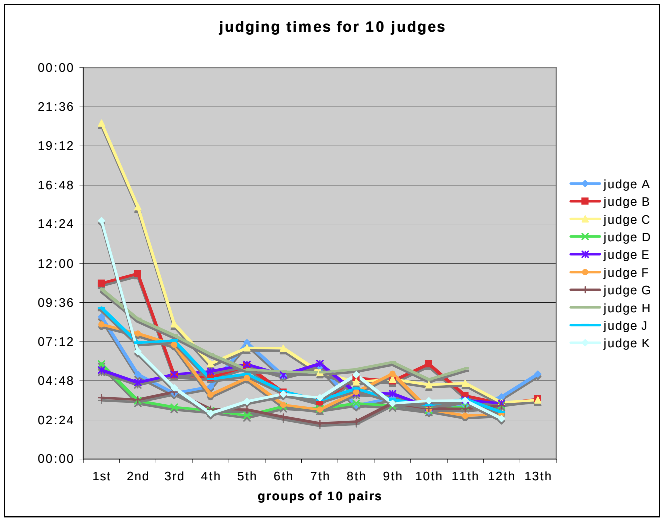
> 16 [of 352] portfolios exceeded this level, [the acceptable ‘fit’ statistic] or 4.5%, which is satisfactory for a 5% significance test.

Concerning Pollitt’s summary of the process

> 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.
>
> (All quotations taken from Pollitt’s report to TERU published within Kimbell et al 2009. All comments in square brackets are our additions for clarification)

e) the scalability of judging

It is a matter of some interest – not least for Awarding Bodies - how long it takes to do the judging. Is this really a process that could be scaled up to become part of a national system of assessment? The pairs engine automatically collects data on the judging process (eg the time taken for each judgement) and these timings – along with the comments of the judging team - are illuminating.



We monitored the time taken by judges to complete groups of 10 paired judgements. The first group of ten takes longest as the judge is coming to terms both with the portfolios and with the pairs engine interface. The 2nd group is typically quicker and the third quicker again. By the 3rd or 4th group, judges have typically reduced their judgement time to between 3 and 6 minutes, and by the 8th group they have reduced it further to between 2 and 5 minutes. The median time for making a paired judgement – across the whole of the 130 judgements and across the whole judging team – was 4 minutes 6 seconds.

This means that each judges’ allocation of 130 paired judgements took them approx 8.5 hours. Based on our work in phase 2, we had estimated 10 hours, but the quicker process in phase 3 is undoubtedly explained by the efficacy of the new pairs engine. The scalability question is how this compares to the kinds of GCSE coursework assessment that teachers are doing currently.

Asked to compare the judging process with his experience of conventional GCSE marking and moderation, one of our judges (an ‘advanced skills’ teacher and head of department) commented as follows:

> With conventional GCSE portfolios it has, in the past, been quite a "painful" experience doing the marking. Usually the first few can take up to an hour each for the larger (better??) ones and reducing down to about 20 or 25 minutes as I "tune in" to the marking criteria. I have also spent quite a time (an hour or two) pre-reading a few folios to get a feel for their overall standard and a rough rank order. Of course added to this can be a few hours (3-4) of internal cross moderation when there is more than one specialist option or more than one teacher marking work from the same exam board.
> A group of 20 folios including internal moderation and administration can therefore easily take 15+ hours... Ok, I am quite methodical, but I do have quite a lot of experience as well!
>
> As to which I prefer.... No contest! E-scape judgements win hands down. The time taken is dramatically reduced for the marking; there is no further administration to do or internal/external moderation. I would also have the added benefit of seeing what has been produced by other schools, something normally only available to examiners and moderators... a great bit of CPD!
> (in Kimbell et al 2009 p 191)

E-scape next steps

A number of initiatives are now underway making use of the e-scape system in whole or in part. These initiatives have been aided by a technical development of the system that now enables e-scape to run from a USB stick – thereby making it possible for schools to use their existing digital kit. Phase 3 of e-scape was based on particular mobile phone/pda devices, but from now on schools can use any lap-top / desktop / camera / mobile that has USB connectivity, and it is entirely cross-platform (mac / pc / linux). If learners’ devices are web connected and logged on, their work saves automatically to their web-portfolio. If they are not, then it saves to their USB stick and the next time it is inserted to a web-connected device it automatically syncs to their portfolio. These developments are now enabling a range of development projects of which the following are representative.

In Scotland, in association with Edinburgh University, Learning & Teaching Scotland and the Scottish Qualifications Authority, a project is underway to explore the *formative* assessment value of the system with learners in the primary/secondary transition years.

With Cambridge International Examinations, we are developing an English: Speaking & Listening component of their IGCSE. This has been through a series of school trials since Jan 2009 and will become a ‘live’ pilot examination component from Sept 2009.

In Western Australia, in association with Edith Cowan University and the WA Curriculum Council an ARC linkage project is underway exploring the use of e-scape in school-leaving examinations in Applied IT, Italian, Physical Education Studies and Engineering. This 3 year project is currently in its 2nd year, with school trials scheduled for Sept/Oct 2009. (see for example http://csalt.education.ecu.edu.au/downloads/AIT_Report2008.pdf)

In England, the next evolution of e-scape is to take it into an Awarding Body pilot that will involve national awards at 16+. This is planned to begin in Sept 2009 in association with at least one Awarding Body and run for the two years of normal 16+ examination courses.

References and related bibliography

For e-scape research reports and other materials please go to: http://www.gold.ac.uk/teru/projectinfo/

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

Kimbell RA, Wheeler A, Sheppard T, Brown-Martin G, Perry D, Hall P, Wharf W, Mller S, Potter J 2005 e-scape portfolio assessment (e-solutions for creative assessment in portfolio environments) phase 1 report pp 68 TERU Goldsmiths University of London

Kimbell, RA Wheeler A, Miller S, and Pollitt A. 2007 e-scape portfolio assessment (e-solutions for creative assessment in portfolio environments) phase 2 report pp 100 TERU Goldsmiths University of London ISBN 978-1-904158-79-0

Kimbell RA (2008 [i]) "Project e-scape: a web-based approach to design and technology learning and assessment" ch 12 (pp219-241) in The episteme Reviews: Research Trends in Science, Technology and Mathematics Education. Eds Choksi B and Natarajan C,. Macmillan India Ltd. New Delhi. ISBN 10:0230-63443-5 ISBN 13: 978-0230-63443-5

Kimbell RA (2008 [ii]) "Design Performance: DigitaL Tools: Research Processes" in Middleton H (eds). Research Methods for Technology Education, Rotterdam Sense Publishers

Kimbell RA and Pollitt A (2008) "Coursework assessment in high stakes examinations: authenticity, creativity, reliability" Third international Rasch measurement conference. Perth: Western Australia: 22nd-24th Jan 2008

Kimbell R, Wheeler T, Stables K, Sheppard T, Martin F, Davies D, Pollitt A, Whitehouse G (2009) e-scape portfolio assessment: phase 3 report. pp 168. Technology Education Research Unit, Goldsmiths University of London.

Pollitt A (2004) "Let's stop marking exams" a paper presented at IAEA Conference, Philadelphia, June 2004.

Pollitt A and Crisp V. (2004) "Could Comparative Judgements Of Script Quality Replace Traditional Marking And Improve The Validity Of Exam Questions?" A paper presented at the British Educational Research Association Annual Conference, UMIST, Manchester, September 2004.

Shore C and Wright S (1999) "Audit Culture and Anthropology: Neo Liberalism in British Higher Education" Journal of the Royal Anthropological Insitute, 5 (4) (December): pp 557-575.

Stables, K. (2008). Embedding digital tools in creativity activity: supporting and assessing the development of creativity'. In J.-C. Hong & Y.-F. Pan (Eds.), International Conference on Creativity Development (pp. 33-52). Taipei: National Taiwan Normal University

Tayler M & Hallgarten J 2000 "Freedom to Modernise" in Education Futures Published in London by a collaboration between the Design Council & Royal Society of Arts.

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