Cambridge Exam Research

www.camexam.co.uk

Abolishing marksism and rescuing validity

Alastair Pollitt

Cambridge Exam Research

A paper for the 35th Annual Conference of the

International Association for Educational Assessment

Brisbane, September 2009

Abolishing marksism and rescuing validity

Abstract

What could be more valid than judging that one piece of work is more creative than another? Or more effective? Or just better? And if many judges agree that the same one is better, isn't that the best evidence for validity we could ask for?

This paper describes progress in applying comparative judgement (first reported to IAEA in 2004) to the assessment of holistic traits like overall achievement, including effectivenes, quality and creativity.

Marking was invented (in Cambridge) during the 18th century enlightment, not in pursuit of validity or even reliability but to overcome serious problems of bias and prejudice in the examinations of the day. Its unintended consequence has been a most serious loss of validity in most of our formal assessments.

Some progress has been made in abolishing marksism in UK assessment. A web-based system has been developed for presenting pairs of 'scripts' and collecting judgements, and the estimation procedure has been shown to be remarkably robust with the extremely sparse data 'matrices' that result. A simple initiation algorithm has been used developed. Some technical aspects of the procedure are described, and a procedure for qualitative description of the scalel for public use id described.

Introduction

Five years ago I presented a paper at the IAEA conference in Philadelphia which argued for the use of paired comparison methodology as an alternative to marking examination papers (Pollitt, 2004). I had introduced this technique to the UK examinations business in 1995 as a technique for monitoring the comparability of different examinations that were meant to share a common standard.

Since 1995 I have become convinced that it has a wider role to play, and the 2004 paper began the process of exploring its applicability, not to comparability studies, but to general assessment. In this paper I will argue that the method can be used to guarantee validity for some kinds of assessment, and to improve validity in many more. I will also report on progress towards the goal of abolishing this unfortunate practice that we have bogged ourselves down in for two centuries.

There are two main theses that this paper presents:

1 The invention of marking has deflected assessors away from their proper focus on validity.

2 In many assessment contexts, paired comparison methodology will enable us to return to concentrating directly on validity.

The origin and consequences of marks

Marking is said to have been invented, in Cambridge, in 1792 or 1793 (Haley & Wothers, 2005; see Stray, 2001, for a fuller discussion). Written examinations were introduced in 1680, and it is a wonder that it took so long to resort to numbers: the 'Tripos' examination in Cambridge University around that time was a daunting affair (Hilken,1967). The examination was principally mathematics, with a small component of theology, and set in three tiers called Wranglers, Optimes and the Polloi. In general the Optimes and Polloi answered questions dictated to them as fast as the fastest students could handle them:

> *It requires everyone to use the utmost dispatch; for a soon as ever the Examiners perceive any one to have finished his paper and subscribed has name to it, another Question is immediately given.* (Wordsworth, 1877, p46)

A Wrangler was given a printed paper of "problems" to take away to complete as much of as possible at "any window he pleases", where paper and ink were set out. The lower classes completed four papers each day, from 8 till 5, and the Wranglers sat an extra two hour paper in the evening. At 5pm on the fourth day, when everyone had done sixteen or nineteen papers, the exam ended, and evaluation began. There were typically about 12 examiners, and after considerable discussion and debate, at midnight the final results and rank orders, including division into eight classes of degree, were posted.

As the number of candidates rose from a handful through dozens to hundreds by the nineteenth century it was clear that some 'technology' was needed. William Farish, Professor of Chemistry then Engineering, introduced the use of marks during his tenure as Proctor of examinations in 1792 and 1793. But there was another reason for this introduction.

Richard Watson graduated in 1759. He remained in Cambridge and became Regius Professor of Divinity until he moved in 1782 to be Bishop of Llandaff (Cardiff). In his autobiographical 'Anecdotes' he wrote:

> *I was the Second Wrangler of my year, the leading moderator having made a person of his own college, and one of his private pupils the first, in direct opposition to the general sense of the examiners in the Senate House, who declared in my favour. The injustice which was done to me then was remembered as long as I lived in the University, and the talk about it didi me more service that if I had been made Senior Wrangler.* *(Watson, R ,1818)*

When a dozen mathematicians are faced with a next to impossible information processing task, it is unlikely that they will not turn to numbers for help. Farish's innovation, it seems, was to add up these numbers across all twelve examiners to avoid serious unfairness of the kind reported by the Bishop of Llandaff – a defence against bias and prejudice.

Single marking rather than multiple

Sixty years later, as Galton (1869) rerported, candidates were still sitting many papers, and the maximum possible mark was around 17,000, implying many markers:

> *The examination lasts five and a half hours a day for eight days. All the answers are carefully marked by the examiners, who add up the marks at the end and range the candidates in strict order of merit. The firmess and thoroughness of Cambridge examinations have never had a breath of suspicion cast upon them.*
>
> *… the marks are not published. They are not even assigned on a uniform system, since each examiner is permitted to employ his own scale of marks.* *(Galton, 1869, p 15)*

Then, perhaps, laziness set in as the practice spread. Not everyone was prepared to countenance the use of twelve markers, and attention turned from assessing the quality of the candidates' work to making do with less effort.

Reliability

Since those days we have come to assume that 'reliable marking' is a pre-requisite for validity, something that was clearly not required in the Cambridge Tripos. The focus shifted from the performance to the question, moving away from both the candidates and the overall quality towards the individual items used to elicit a part of that performance. In this shift, validity disappeared.

We began to concentrate on writing questions that lend themselves to reliable marking, and especially to multiple choice, to minimise the role of judgement in marking - rather than on writing questions that will be intrinsically valid by eliciting genuine evidence of the things we want the candidates' to show us they can do (Pollitt & Ahmed, 2001; Pollitt et al, 2008).

The current UK examination system imagines an ideal scenario in which a single marker marks all of the thousands of papers in an exam, to a constant standard. Since this is not practicable Assistant Markers are employed to help, but they are not supposed to think, and should simply behave as if they were ‘clones’ of the Principal Marker.

Thus professionals are hired to act in a way that suppresses their professionalism, while marking questions that have been designed to allow this automated process rather than to chieve valid assessment.

The story of validity

The current orthodoxy on validity is not helpful to test constructors.

> *Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. … It is the interpretations of test scores that are evaluated, not the test itself.*
>
> *(AERA et al, 1999, p9)*

There is a serious danger with this view, that validity will be seen as the business of test interpreters, *rather than* of test constructors. There would seem to be little point in a question writer trying very hard to make better tests if the users are going to misuse them wantonly. The second chapter of *Standards* is titled *Reliability and errors of measurement*, and at least 18 of its 20 “Standard” statements are aimed more at test constructors than at test users. Test constructors need a model of validity that is relevant to their concerns, rather than being urged merely to maximise reliability.

It is folly even to suggest that validity is a concern only when results are being interpreted; validity *cannot* be present in the interpretation if it was not built into the test from the beginning. We have long argued for a concept of *intrinsic validity,* (Pollitt & Ahmed, 1999) where the primary responsibility for validity lies with the people who design the test and those who write the questions: if they put garbage in, no one further down the line can deliver anything other than garbage out. Recently, Borsboom (2005) made a similar argument from the perspective of the philosophy of science: “*the* [integrated validity] *theory fails to serve either the theoretically oriented psychologist or the practically inclined tester*” (p150). He argues that validity depends on the existence of a *causative* link from the trait being measured, through the items in the test, to the resulting scores:

> *Validity is a property of tests: a valid test can convey the effect of variation in the attribute we intend to measure.* *(p162)*

We would add that it is also a property of each and every question in the test, and that this conveying of the effect of the attribute can continue through administration, scoring and reporting.

To summarise, test constructors need a model where validity is a continuous quantity, a property of the assessment process, maximised at the beginning – when the test is conceived – and lost to some degree at every step along the way.

The paired comparison method and validity

Intrinsic validity is effectively guaranteed by the paired comparison method, if it is properly applied. Because validity is intrinsic to the procedure, as judges judge directly against the statement of what matters (see below), the amount of validity is *not* limited by unreliability, since the validity is intrinsic to the procedure. If you want to conceptualise reliability in such a system you will find instead that the reliability is constrained by the validity – but there is no need to conceptualise reliability when more important measures of accuracy and standard error are more directly available.

Method

The method is fully described in Pollitt (2004) and reviewed in detail in Bramley (2007).The initial presentation is in Thurstone (1927a; 1927b).

Progress

Since 2004 several issues have been explored, and some have been resolved. The most significant test-bed has been the **e-scape** project (TERU, 2007), which is also described in Kimbell (2009). In this, students compile an electronic portfolio under controlled conditions, with about 30 pages containing text, drawings, photographs, and audio recordings, responses to prompts, notes and conclusions and reflections. A series of studies have established that paired comparison is capable of providing measures of the quality of the students' work that correlate highly with other measures and are more 'reliable' than marking could deliver in with the same time and effort. Here are some of the steps we have taken; in these notes the word 'script' will be used to stand for any piece of evidence about a student's performance, not necessarily written on paper.

1 The issue of what instruction to give judges before they begin has been clarified. The TERU work has shown that a simple comprehensive statement expressing what is *Important* in the subject being assessed is a fully satisfactory definition of the task; a short discussion to make sure every judge understands what is and isn't important is the only training needed – and this does not need to be repeated for each task.

As a starting point we use the *Importance* statements produced for English schools by the QCA (for an example see QCA, 2009). These are brief enough to be memorised by subject specialists, and constitute a consensus view of what the teaching of the subject will emphasise – and what the assessment will value. We see these statements, or statements like them, as an essential starting point for developing any examination – rather than any list of aims or objectives, or any taxonomy of behaviours, these statements tell everyone involved in the teaching and assessment what matters, in simple language that they can use to design validity into their part of the programme (Pollitt et al, 2008).

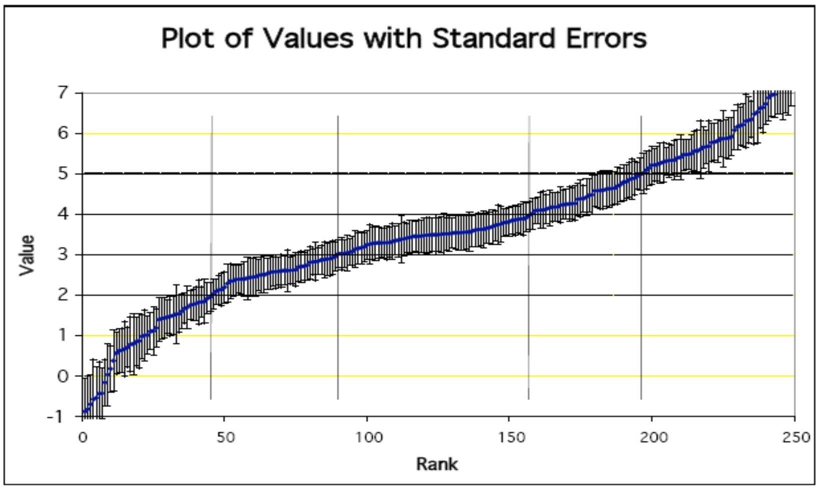
2 An interactive web-based system has been developed for collecting performances; it is suitable for any form of product, or record of process, that can be captured electronically. The system is so far attracting most interest for *performance* and *portfolio* assessment, using electronic objects, audio or video, but it is easily applied to scanned paper performances too. The system can run an assessment exercise with a (so far) unlimited number of judges making paired comparisons in either a chained or an unchained sequence (see below), and can calibrate and re-calibrate the judgements so far as frequently as desired.[1]

3 A design for calibration and for optimised selection of the next partner for comparison has been implemented that produces stable estimates of the quality parameter for each piece of work more quickly that we thought possible in 2004.

Several systems have been designed for initialising the calibration; currently we use a 'Swiss tournament' method, based on practice in chess tournaments to set up the first six rounds, before the optimised system takes over. Other options can be implemented too.

[1] Credit for this is due to Declan Lynch and Karim Derrick of TAG Learning, partners in the e-scape project .

4 Standard setting has been designed into the system, using a selected subset of judges – either 'experts' or judges who are necessarily unbiased – to compare new with old performances. This diagram shows the result of standard setting as part of the system.



The horizontal lines show grade boundaries set by judgement. Extra judgements have been targeted at scripts which were found to be near the boundaries, so automatically increasing the classification consistency of the procedure.

5 The possibility of asking judges to rank order a small set of scripts, rather than simply compare two, has been considered, and will be explored once enough experience of paired comparison has been collected. Bramley (2007) outlines the possible advantages of this approach.

6 The e-scape scheme, and two other exam components using the same assessment engine, are due to go live this autumn, as full examination modules, in three GCSE subjects. Some other certification applications, notably in language testing, are being explored.

7 A university in Ireland is using the system for internal assessment, and for peer assessment. Asking students to judge between the work of their peers in this way is a new development in education, and promotes something we have advocated for years – if you can teach students how to evaluate their own work you have taught them how to do it well themselves. (See Dimitrova, 1996, for an illustration of this in the teaching of writing.)

Recent experience

Two critical issues have been explored through the series of e–scape trials. We were concerned that the system could not scale up, because the theoretical basis of Thurstone's model would break down when the matrix of possible comparisons became too large, and we suspected that the chaining procedure used in many studies might introduce a bias that would invalidate the outcomes. This section reports on these two issues.

1 Size

In a typical comparability study, as carried out from 1977 onwards, 30 scripts from 6 boards would be judged. Since only comparisons between boards were made, there were (30*25)/2 or 375 possible comparisons: typically, about 1500 judgements were made, giving about 4 judgements for each possible comparison.

The basis of analysis according to Thurstone's model is that the log-ratio of wins to losses for each possible pair estimates the difference between them on the quality scale. If there are just 4 comparisons, then there are only 3 possible estimates:

log(3:1) => 1.10, log(2:2) => 0.00, log(1:3) => -1.10

Such a restricted set of data values must seem a very dubious basis for estimating measures of script quality. Yet it works – with considerable accuracy, because of the richness of the inter-linking between the large number of pairs involved. Jones et al (2004) ran one of these studies in duplicate, using two independent sets of judges, and obtained correlations above 0.8 betweens the two sets of parameters; given the extremely restricted range of quality in the scripts around a single GCSE boundary, this is a very high reliability coefficient.

To replace marking with paired comparison, however, we cannot afford so many judgements for every possible comparison. To score 100 scripts in a similar manner would need 4*(100*99)/2 = 19,800 judgements: in practice this would mean almost as many judges as students. To deal with 1000 students would need 2000 judges! An obvious question, then, is how few judgements we need.

To explore the question we first simulated trials using subsets of existing data from various studies. It was quickly clear that replicated analyses with an average of just one judgement for each possible comparison were still reliable. We then turned to new data, collected in the **e-scape** project. In the first large pilot (called e-scape 249) we used a manually operated, partially targeted, system for selecting comparisons. There were 249 scripts from the whole ability range and each was judged, on average, 18.6 times. The table below shows some pertinent figures:

	Study, N scripts	Notional Matrix size	Nj = number of judgements	N j per script	Matrix: average cell entry	alpha coeff
1	Comp 30	810	1,500	60	5.00	
2	e-scape 249	30,876	2,322	18.6	0.075	0.93
3	e-scape 352	61,776	3,097	19.9	0.050	0.95

In this study the matrix was very empty: only about 7.5% of the cells contained anything at all, and that was nearly always just a 1 or a 0. Only by chance did any comparison get replicated by a second judge. Yet the analysis held up, and the resulting parameter values made sense. The alpha-like measure of internal consistency amongst the parameter values was 0.93, a figure at least as high as most GCSE examination components could achieve, and almost certainly higher than any other judgementally scored component. In a blind comparison study, 20 of these scripts were marked by several markers, using a traditional mark scheme, and the rank order correlation between the marks and the parameters calculated. The value of 0.88 was high enough to confirm that the paired comparison method was valid enough for its purpose.

We noted that the number of comparisons needed can be substantially reduced by realising that there is little point in comparing scripts that are far apart in quality, as the difference between them is obvious and not very informative. In statistical terms, information equals *p* times (*1-p*) where p is the probability that one script will beat the other. If *p* is below 0.3 or above 0.7 the amount of information gained falls off fast. The **e-scape 352** trial was fully automated. Parameters were re-estimated approximately every time each script had received an extra judgement, and an optimal partner was chosen for its next comparison. The system more or less reached stability after only 12 judgements per script, but we continued to 15 each; after that, extra comparisons were added for particular scripts to simulate estimation around boundaries (see Progress 4 above). The scale consistency was very high, with an alpha-like coefficient above 0.95 for the full range of quality we would expect in a GCSE examination, yet the total time spent by the judges was estimated to be less than the time they would have taken to mark the scripts using the traditional mark scheme.

There is no sign so far that the analysis breaks down when the matrix is very sparse. What appears to matter is the *local linkage* of scripts, that is the number of comparisons between each script and its near neighbours. It seems that if every script is compared to at least 12 other scripts where the probability of a win is between 30 and 70% (and, of course, that the whole set of scripts are 'connected' somehow) then the analysis will succeed in creating a consistent scale. The data are better imagined as a rope of inter-twisted strands, where it is the local strength that determines the overall strength, rather than as a network or web: as a consequence the nuber of judgements needed for the system to work is approximately a linear function of the number of scripts to be evaluated.

2 Chaining & Bias

One feature of the **e-scape** studies was that most of the comparitive judgements made were 'chained'. Chaining judgements means that after comparing two scripts A and B, a judge then compare B and C, then C and D, and so on. This strategy was introduced early in the use of paired comparison in the comparability studies for efficiency, to reduce the time a judge needs to spend becoming familiar with the two sets of work before making a decision. There was, of course, some concern that this might result in bias:

> *However, it does have the drawback of probably violating one of the assumptions of the Thurstone pairs method – that each paired comparison is independent of the others. If the same script is involved in consecutive pairs of judgments then it is highly likely that features of it will be remembered from one comparison to the next.* (Bramley, 2007, p266)

So far, no one who has carried out a comparability study has detected this kind of bias, but it did seem more likely than not that it should happen. In traditional marking, a similar kind of effect, often called a halo effect, has frequently been found in which scores by the same student on different questions correlate more highly when they are marked by the same marker than when they are marked by different markers.

In a scoring context, the amount of work to be read is usually less than in a comparability study, since only one component is judged instead of the whole examination. Nevertheless, it still seemed advisable to use chaining to maximise efficiency, and a study was designed to search for the expected bias effect.

The basis of such an analysis was outlined in Pollitt & Elliott (2003). Each time a judgement is made, it can be scored as '1' if the first script wins or '0' if the second wins. The probability of the first one winning can be calculated from the final quality parameters:

$$\text{odds (first wins)} = \exp(\text{parameter}_1 - \text{parameter}_2), \text{ and}$$
$$\text{probability (first wins)} = \text{odds}/ (\text{odds} + 1)$$

This probability will always be a number *between* 1 and 0, never equal to either. This means there will always be a *residual*, the difference between the 'observe' score of 1 or 0 and the 'expected' score or probability. These residuals can be standardised, summed and otherwise manipulated to give chi-square tests of hypotheses about the fit of the data to the ideal model.

As a first step, the outcomes – wins and losses – from the **e-scape 352** study were explored. By finding the average parameter values for every comparison we can predict the number of wins in each of several circumstances, and test the observed number against this expectation.

Average:	**All**	**O**	**E**	**χ2**	**p**
N 1st wins =		1464	1532.40	6.040	0.014
N 2nd wins =		1591	1522.60		
Average:	**Chain**				
N 1st wins =		964	1019.18	5.930	0.015
N 2nd wins =		1055	999.82		
Average:	**No chain**				
N 1st wins =		500	513.23	0.630	0.427
N 2nd wins =		536	522.77		

For all the data, there were fewer wins by the first script than expected (p = 0.014). When we split the data into ‘Chain’ or ‘No chain’, there is clearly no effect from comparisons that were not in chains. There were fewer wins by first scripts only in the chained data. First scripts are the ones that have already been judged in the previous comparison. To explore why this happens, we need to split the data further, to separate comparisons in which the first script won, or lost, its previous comparison. The table, for chained comparisons only, was:

Average:	**Chain**	**after Winning**			
N 1st wins =		661	667.47	0.140	0.708
N 1st loses =		397	390.53		
Average:	**Chain**	**after Losing**			
N 1st wins =		303	362.95	15.650	< 0.0001
N 1st loses =		658	598.05		

The whole of the misfit seems to be concentrated in the cases where a script has lost its first comparison; it seems then that it is more likely to lose the second comparison than the final parameters predict. There seems to be no such effect if a script wins its first comparison.

A different analysis, using the residuals, comes to a similar conclusion.

	Mean residual	Mean theoretical probability
Win after win	0.370	**0.630**
Win after loss	0.364	**0.636**
Lose after win	-0.417	**0.417**
Lose after loss	-0.379	**0.379**

If we consider only the scripts that win their second comparison, their average theoretical probability of winning does not differ significantly whether they won or lost their first. But in the case of those that lose their second comparison, there is a small difference. The ‘Mean probability’ can be taken as an indication of the average quality of the scripts in that sub-group: *better* scripts are losing after an initial win than are losing after an initial loss.

The difference is small, and limited to this sub-group that lose in their second comparison, having won their first. It is hard to suggest an explanation: perhaps judges who carry forward a winner try subconsciously to counter any halo effect in the second comparison, so that if they decide to award a second win they then ‘double-check’ for safety, but they may not do this for a double loss.

Three things are important:

(i) the effect is very small, and limited to about one sixth of the chained comparisons (nb: not one sixth of the scripts);

(ii) chaining itself makes it unlikely that this will happen often for a given script;

(iii) paired-comparison methodology is more able to detect and, if necessary, to correct for bias effects of this kind than marking has ever managed to do.

We will study this effect in future trials and, if it persists, modify the algorithm that selects comparison pairs to compensate for it.

What are we measuring?

One challenge to the method has been based on the notion of lack of transparency – that there is no mark scheme to show where each script succeeded or failed. One approach to resolving this issue exploits the idea of comparison in a qualitative context.

In the first published educational assessment study using Thurstone's method, Pollitt & Murray (1993) asked judges both to choose between a pair of video recordings of Speaking in a foreign language, and then to describe quickly what they saw as similarities and differences between the two performances. This second part was an implementation of Kelly's *construct elicitation* technique, based on his Personal Construct theory (Kelly, 1955), and it generated a theoretically relevant definition of the scale *as perceived by the judges*. More recently, this element has also been added to the paired comparison procedure in several of the comparability studies referred to earlier (see Pollitt et al, 2007, for some examples).

In any application of paired comparison to high stakes assessment, eliciting a descriptive scale in this way will be a relatively small exercise, much cheaper than employing markers and senior examiners to write reports. The results each year will accumulate into a full principled description of the trait being assessed, an elaboration of the importance statement that began the whole process.

Conclusions

We are progressing, slowly, towards the goal of abolishing marksism.

In reality, the aim is not to abolish marking everywhere but to extend gradually the range of contexts in which it is seen as appropriate to use judgement. Paired comparison is naturally more appropriate than marking for assessing performances and any product which captures evidence of the process which led to it. It may be the best way to assess any form of essay writing, and experience has shown that it *can* (in some circumstances) be used successfully even in assessing in such naturally countable domains as mathematics.

Direct comparative judgement removes the need to design questions for reliable marking, sustituting instead the need to design them to elicit valid evidence of achievement. Where it can be used, it seems very likely to improve not only the accuracy of each student's assessment but the overall validity of the procedure as well.

Paired comparison can bring us back to the 'natural' system that existed in the eighteenth century, in which many judges pooled their evaluations without resorting to the dangerous practice of assigning numbers to the quality they saw, and with greatly improved control over the quality – validity – of the process.

References

AERA, APA, NCME (1999) *Standards for educational and psychological testing.* Washington, AERA.

Borsboom, D (2005) *Measuring the mind*. Cambridge : Cambridge University Press.

Bramley, T (2007) *Paired comparison methods.* In Newton, P, Baird, J, Patrick, H, Goldstein, H, Timms, P and Wood, A (Eds) *Techniques for monitoring the comparability of examination standards*. London, QCA.

Galton, F (1855) *Hereditary genius : an inquiry into its laws and consequences*. London : Macmillan.

Haley, C & Wothers, P (2005) in Archer, MD & Haley, CD (Eds) *The 1702 Chair of Chemistry at Cambridge*. Cambridge, CUP

Hilken, TJN (1967) *Engineering at Cambridge University: 1783-1965*. Cambridge, Cambridge University Press.

History of Engineering at Cambridge University, 1783 - 1965 by T.J.N. Hilken, published by Cambridge University Press in 1967.

Jones, BE, Meadows, M & Al-Bayatti, M (2004) *Report of the inter-awarding body comparability study of GCSE religious studies (full course) summer 2003.* Manchester, AQA.

Kelly, GA (1955): *The Psychology of Personal Constructs*, vols. I and II, Norton, New York.

Kimbell, R (2009) *Developing more effective assessment practices: Constraints, authentic evidence and electronic portfolios.* IAEA, Brisbane, Australia, September 2009.

Pollitt, A (2004) *Let's stop marking exams*. IAEA, Philadelphia, June 2004. Available at http://www.camexam.co.uk/ Our publications.

Pollitt, A, & Ahmed, A (1999) *A New Model of the Question Answering Process*. IAEA, Bled, Slovenia, May 1999. Available at http://www.camexam.co.uk/ Our publications.

Pollitt, A, & Ahmed, A (2001) *Science or Reading?: How students think when answering TIMSS questions*. International Association for Educational Assessment, Rio de Janeiro, Brazil, May 2001. Available at http://www.camexam.co.uk/ Our publications.

Pollitt, A, Ahmed, A, Baird, J-A, Tognolini, J & Davidson, M (2008) *Improving the quality of GCSE assessmen*t. Final report to QCA. Available at http://www.camexam.co.uk/ Our publications.

Pollitt, A, Crisp, V and Ahmed, A (2007) The demands of examination syllabuses and question papers. In Newton, P, Baird, J, Patrick, H, Goldstein, H, Timms, P and Wood, A (Eds) *Techniques for monitoring the comparability of examination standards*. London, QCA.

Pollitt, A & Elliott, G. (2003) *Monitoring and investigating comparability: a proper role for human judgement*. Paper given at the QCA 'Comparability and Standards' seminar, Newport Pagnell, 4th April. Available at http://www.camexam.co.uk/ Our publications.

Pollitt, A, & Murray, NJ (1993) *What raters* really *pay attention to*. Language Testing Research Colloquium, Cambridge. Republished in Milanovic, M & Saville, N (Eds), *Studies in Language Testing 3: Performance Testing, Cognition and Assessment*, , Cambridge: Cambridge University Press. Soon available at http://www.camexam.co.uk/ Our publications.

QCA (2009) The importance of Design and Technology.

http://curriculum.qca.org.uk/key-stages-3-and-4/subjects/design-and-technology/index.aspx

Stray, C (2001) The Shift from Oral to Written Examination: Cambridge and Oxford 1700–1900. *Issues in Educational Assessment : Principles, Policy & Practice,* 8, 33-50.

TERU (2007) ***e-scape*** *e-solutions for creative assessment in portfolio environments*. Goldsmith's College, University of London.

Thurstone, L.L. (1927a). Psychophysical analysis. *American Journal of Psychology, 38,* 368-389. Chapter 2 in Thurstone, L.L. (1959). *The measurement of values.* University of Chicago Press, Chicago, Illinois.

Thurstone, L.L. (1927b). The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology, 21,* 384-400. Chapter 7 in Thurstone, L.L. (1959). *The measurement of values.* University of Chicago Press, Chicago, Illinois.

Watson, R (1818). *Anecdotes of the life of Richard Watson ... written by himself at different intervals, and revised in 1814*. Published by his son, Richard Watson, LL.B., prebendary of Landaff and Wells. London : T. Cadell and W. Davies.

Wordsworth, Christopher, (1877) *Scholae Academicae*. London, Frank Cass.