RESEARCH AND EVALUATION DIVISION

www.ucles-red.cam.ac.uk

Let's stop marking exams

Alastair Pollitt

with assistance from Gill Elliott and Ayesha Ahmed

University of Cambridge Local Examinations Syndicate

Paper presented at the IAEA Conference, Philadelphia, June 2004.

Contact details

Alastair Pollitt, RED, UCLES, 1 Hills Road, Cambridge, CB1 2EU

pollitt.a@ucles.org.uk

UNIVERSITY of CAMBRIDGE
Local Examinations Syndicate

1 Hills Road, Cambridge, CB1 2EU

Let's stop marking exams

Alastair Pollitt

with help from

Gill Elliott and Ayesha Ahmed

April, 2004

Acknowledgement

This paper has grown from a seminar presentation Gill Elliott and I made (Pollitt & Elliott, 2003b) at a QCA meeting concerned with studies of the comparability of standards in different examinations supposedly assessing the same subject. Because of Gill's absence for the past few months, I take full responsibility for the extension of our thinking then, while owing a lot to the foundation she and I created together. I am grateful too for some excellent discussions of the ideas I have had with Ayesha Ahmed, and with other colleagues in Cambridge.

Introduction

In this paper I will propose an alternative method for carrying out summative assessment, one that seems to be intrinsically more valid than the familiar procedure of awarding marks to lots of questions, little or large, and adding them up to get a student's total mark or score.

Although the new approach may seem radical it is rooted in the psychophysics of the 1920s; practical considerations have, however, made it impossible to apply the procedures in this way until now. I shall argue that the method brings us much closer to the essential purpose of summative assessment and that, although many concerns still need to be addressed, there are enormous advantages to be gained.

I begin by considering the quantitative theories that are supposed to underpin the business of certifying educational achievement.

Formal theories for educational assessment

Two phrases are commonly used to refer to the mathematical theories used in assessment – 'measurement theory' and 'test theory'. In fact, they are often used as if they were synonyms, though the former really should be more general then the latter. Browsing a library leads to a rather curious finding: whichever term you begin with, you will very soon find that you are reading about a theory of test *scores*, as if it were necessarily true that tests yield scores.

Confirmation of this finding can be found in the formal general statements about test theories that can be found in textbooks. For example:

> *The purpose of any test theory is to describe how inferences from examinee item responses and/or test scores can be made about unobservable examinee characteristics or traits.* (Hambleton & Swaminathan, 1985)

I have no complaint about this statement of purpose – except for the phrase "item responses and/or test scores". Why must our test theory be about *items* or *scores*? Are there no other possible outcomes from a test, about which we can construct a formal theory?

Traditional test theory

The presumption of scores is most obviously seen in what is nowadays called either *traditional* or *classical* test theory, and which is based on the very simple equation:

$$x_o = x_t + e$$

In this, 'e' refers to the error (assumed random) that will accompany any measurement process and the other two terms refer to the *observed* and *true* scores for a given student. In fact, these x's may refer either to item scores or to test scores, where the test score is the sum of all the item scores, but the important point for us is that the theory is absolutely explicitly about *scores*.

Modern test theory

If we consider more modern test theories, things are only a little better. The most general term for these is Latent Trait Theory[1], of which we can distinguish two traditions, using IRT and Rasch models. The book that first made Item Response Theory widely known was titled *A Statistical Theory of Mental Test Scores*, (Lord & Novick, 1968); once again it is explicitly about *scores*, which is not surprising given the origin of IRT in Lawley's (1943) attempt to free traditional item analysis from the specific characteristics of the samples of students used in pre-testing. In the Rasch tradition, the most influential book has been *Best Test Design* (Wright & Stone, 1980) which turns out to be wholly about making tests from dichotomous items.

Of course, many books have been written about IRT and Rasch models since these seminal ones, but it is still almost universally true that the authors seem to assume that testing generates ***scores***, at the levels both of items and, by summation, of tests.

Measurement theory

But at the most general level we often consider our business to be *measurement* - an appllication of psycho*metrics* – and it should then become clear that measurement theory ought really to be about estimating measures of mental traits, with no necessary reason at all why we have to assume that these measures will be derived from scores.

Is there an alternative approach to measurement that we have been ignoring? I argue that there is, and to introduce it I return again to a consideration of our business.

Summative assessment

For examination boards, and often for other assessment agencies, the core business is what is described as *summative assessment*, or the assigning to each of a large number of students a number which represents their level of performance on tasks which are designed to discover their level of achievement in some educational area. We may, for various reasons, report that number as

---

[1] It is this notion of 'latent traits', of course, that Hambleton and Swaminathan are alluding to in the quotation above.
For present purposes we can consider Latent Class Theory as a member of the group of latent trait theories.

a letter grade or even merely as a 'pass' or 'failure' but this does not alter the essence of the process. We are required to do two things: to sort the candidates into a rank order with sufficient precision and categorisation to meet the needs that our national educational, economic and political systems place on the examination system, and (usually) to attach constant standards to that ordering. The first requirement ensures that those who will use the results to select some students rather than others are given enough information for their purpose; the second provides a system for interpreting the results, for giving meaningful 'reference' to each point on the scale and for monitoring the standards of students' achievement over time and over different examinations.

If we summarise these requirements in a single simple sentence it will read something like this:

THE FUNDAMENTAL PURPOSE OF SUMMATIVE ASSESSMENT:

> *We are required to judge the overall quality of students (or their performances) in some educational domain on a standard ordinal scale.*

There is some difference between countries about whether it is performances or students that are to be ordered. In most, including England, the task of assessment agencies is to sort *performances* into order and these performances are obtained through the process called examination or testing. Occasionally it is possible for appeals against results to lead to the exam result being over-ruled by evidence that the *student* usually performs better than they did in the exam. In a few countries test results are only used to guide the teachers, who are assumed to have the best view of students'achievement.

Although the Fundamental Purpose only requires ordinal measurement it is common for the public, and the most official or sophisticated users of summative assessment to assume that the results are expressed on an interval scale, and it would therefore by desirable for our procedures to generate scales that can support these interpretations.

Why do we mark exams?

Given the Fundamental Purpose, it is not obviously necessary for us to mark exams. The requirement is that we find some way to ***judge*** the students' performances in order to create the scale we need, and marking items to add up their scores is just the way we seem to have chosen to do this. Those countries that depend on school reports rather than exam or test results may avoid marking performances, though very often it seems they use tests as part of the process by which teachers generate their judgements.

There seems to be, quite reasonably, a great deal of concern about the perceived subjectivity of judges and the possible unfairness that might arise in a system with no 'objective' control over the judgement process. The most extreme cases of this, of course, have led to the use of multiple-choice testing but I argue that the same concerns, buttressed by 'traditional test theory', have led many other assessment agencies down the road of miniaturising the elements of an examination. It is assumed (and indeed it is provable) that examiners will differ less in the score they give to a particular performance when that performance is based on many elements each of which is only 'worth' one or a few marks, rather than on a few elements each worth many marks.

In essence we ask our judges to make many micro-judgements and score them so that we can then use simple addition to generate a total score which is used as the macro-judgement required by the Fundamental Purpose. I think this trick is dangerous and that several harmful consequences are likely to follow. Driven by 'traditional test theory' the concept of *reliability* tends to become dominant[2] at the expense of validity. It is easier to challenge or to defend the reliability of a test than its validity, since validity is difficult and expensive to quantify while 'internal consistency' gives a quick and cheap estimate of one form of reliability, albeit far from the most informative one for evaluating educational achievement tests.

Marking is expensive, at least in the English system, where only experienced teachers are considered skilled enough to make the micro-judgements reliably, and it is generally held that full double-marking would be unacceptably expensive. This of course simply increases the pressure to make the micro-judgements as 'marker proof' as possible, and the test questions smaller still. Most seriously, perhaps, the question writers are obliged to write questions that can be marked reliably. Since there may be as many as a few hundred markers for some big examinations, it is obviously a serious constraint for the writers that the questions they set must be capable of being marked by markers essentially acting like automata; they may be unable to ask the question they wanted to ask and forced to distort it for the sake of reliable marking. When examiners are prevented from asking the questions they want to ask, it seems to me that some form of invalidity is almost guaranteed.

Furthermore, why should we expect any weighted summation of micro-judgements to lead to the 'correct' macro-judgement? Given the well known complexities of weighting – the subtleties of intended and achieved weights in a composite score – it seems most unlikely that, just by chance, a total test score should happen to give the optimal measure of a student's performance or ability.

In these circumstances, why do we not ask our examiners to make the macro-judgement directly? Reference to the literature does not easily answer this question. It is difficult to find any serious discussion of judgement, in this sense, in any of the well known textbooks. The third edition of what is perhaps the most respected of all American reference books, *Educational Measurement* (Linn, 1993) refers to "judgement" only in the following senses: in Messick's well known chapter on validity are references to judgement of *content relevance, test content* or *domain content*, of *format*, of *scoring models* or *administrative procedures* and of *measurement contexts*; in other chapters are references to judgements in *standard setting* and *test specification*. In summary, all of the references to judgement are to judging either items or procedures, and never to judging students or performances. A similar story can be told from the British literature. In the survey of *Assessment and Testing* commissioned by UCLES (Wood, 1991) there is no reference to judging except in the quite specific contexts of standard setting and essay marking.

Can we find a theoretical basis for direct judgement as an alternative to marking examination performances?

---

[2] Indeed traditional test theory is sometimes even described as the theory of reliability.

Alternative 1: direct macro-judgement

In some domains, in what is often called *performance assessment*, there is a considerable tradition of rating performances against verbal descriptions of standards. *Grade descriptors* are commonly used in assessing essays in first and foreign language assessment, in assessing speaking ability in FL tests, and in other domains where ratable performances are observed, such as Physical Education, sport, visual art, or drama. This kind of approach has been extended to many vocational areas, and can be used in any domain in which success in a course can reasonably be evaluated by observing a performance or a product that embodies the aims of the course.

There is, though, a serious problem that prevents this approach from being the basis for a fully generalisable approach to summative assessment. It arises from a fundamental difficulty with the nature of judgement. In the words of a recent book on the psychology of judgement:

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

In other words, all judgements are relative. 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. What is the imagined performance that properly embodies a particular verbal descriptor? We need somehow to standardise this prototype in the minds of all our assessors if we are to achieve reliable direct judgement. The difficulty of achieving this is the cause of the well known problems with direct ratings.

First, the result is generally a rather crude scale, with only about five categories able to be reliably distinguished (Laming, 2004, p17). In English examinations it is therefore probably no coincidence that results are reported in terms of five pass grades (A Level or most of the EFL examinations) or eight awarded in (usually) two separate but overlapping tiers of test (GCSE). Nevertheless there is considerable concern about the level of disagreement between raters, such that "inter-marker reliability" dominates technical discussions about the quality of the assessments. In general it is accepted that it is preferable to average the ratings of two raters, or three raters, or more, but considerations of cost quickly intervene.

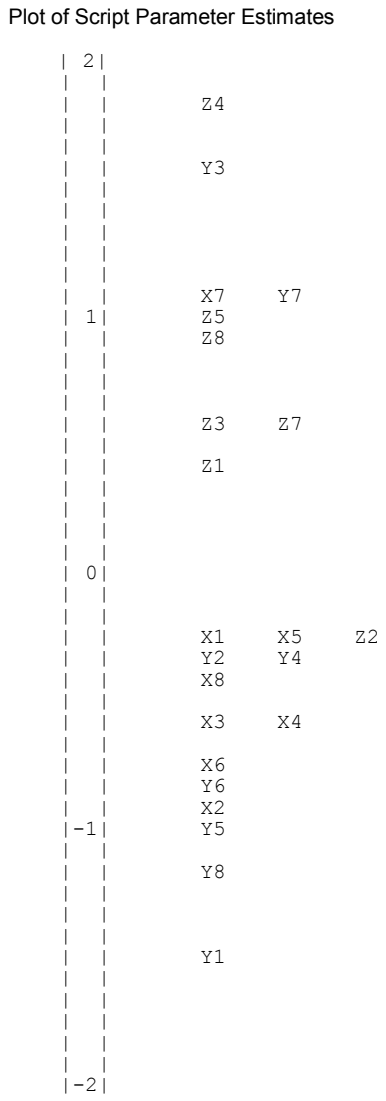
But if *All judgments are comparisons of one thing with another*, why do we not compare performances **directly**?

Alternative 2 direct comparative judgement

The alternative approach to summative assessment that I would like to propose is based on the psychophysical research of Louis L. Thurstone, and specifically on his *Law of Comparative Judgement* (Thurstone, 1927). In terms of the principles developed here, the significance of the approach is that it is one of his methods for constructing scales from human judgement, and is derived directly from logical principles of measurement. The essential point will be familiar to anyone grounded in the principles of Rasch models: when a judge compares two performances (using their own personal 'standard' or internalised

criteria) **the judge's standard cancels out**. In theory the same relative judgement is expected from any well-behaved judge. 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*. In exactly the same way that item banking creates an interval scale measuring the relative difficulty of items or the relative ability of students, so comparative judgement of performances on the same task can construct a true measurement scale expressing the relative value of the performances, as shown on the right.



This figure presents the results graphically, with scripts judged of high quality at the top, and those judged less good lower down. A high quality script has been judged better than most of those below it, and much better than those far below it.

Of course, the analysis will also generate a more detailed table giving two important values: the parameter estimating the quality of each script (in units called *logits*), and a standard error for that estimate. We will return to the importance of that standard error later.

Beyond the familiar properties that comparative judgement shares with item banking are possibilities based on the human ability to make sophisticated adjustments for differences between tasks, and these possibilities will be discussed later. For the moment let us consider how this approach might work in simple summative educational assessment.

How to create a judgement scale (instead of marking)

Step 1

Scripts are sent to judges in pairs - judges report which one is the 'better'

For convenience in the discussion ahead I will use the word *script* in general to stand for any object that a student submits as evidence for achievement, whether it be written, drawn, constructed, video recorded or whatever.

Comparative judgement requires that scripts are sent to judges in pairs, and that the judges simply report which one is the 'better' in each pair. While current 'marking' approaches require only that each script be scored once, comparative judgement will need each script to be seen several times in different pairings. In the circumstances in which major summative asssessment happens in England this is simply not feasible, mostly because there is not enough time between the collection of the scripts (the examination date) and the date on which results must be published. Very soon, however, it is clear that most examination scripts will be digitally scanned, and it will be technologically easy for the examination board to send scanned images of the scripts for quick, even simultaneous, on-screen judgement around the country, or even the world.

Thus Step 1 becomes:

Script images are sent to judges in pairs - judges report which one is the 'better'

Step 2

Early analysis optimises data collection

An important principle in the analysis of comparative judgement data is that every judgement is statistically independent. Experience so far suggests that this requirement is quite robust, and data collection designs that appear to violate the principle in reality show no statistical evidence of doing so. It follows, if this assumption is met, that

> *at any time, data collected so far can be analysed to create a temporary scale*[3]

The importance of this is that it enables early analysis of the data collected so far to be used to optimise the collection of further data. This can happen in two ways; one applies to cases where more information is needed for particular reasons, and will be discussed later, but the other optimisation principle is fundamental to an efficient system.

Information theory tells us that the information we get from a comparison will be maximised if the 'true' probability of one script 'beating' the other is 0.5, that is, when the two scripts are equal in standard. As the difference between the two scripts increases, the amount of information contained in a judgement falls off slowly at first, being still 96% of the maximum when the probability has moved to 0.6 or 0.4, but accelerates downwards as the scripts become less equal, to 64% at probabilities of 0.8 or 0.2, and only 36% at 0.9 or 0.1. These percentages can be interpreted directly as measures of the efficiency of the data collection process, meaning that to maximise efficiency we need to ensure, as early as possible as possible, that we are only asking judges to compare scripts that are closely matched.

[3] There is also a practical requirement that all of the data are 'linked' via overlapping script comparisons.

The table on the right gives an indication of how close the scripts should be in logits to maintain a given level of efficiency. A reasonable target would to aim to ensure the average difference between scripts being compared is about 1 logit, keeping efficiency around 80%, and not more than 1.5.

Efficiency (%)	Parameter difference
100	0.00
90	0.65
78.5	1.00
75	1.10
66.7	1.32
50	1.76

Early analysis of the data from the first judgements will clearly help by giving initial estimates of the parameter values, and these need not be very accurate to provide substantial improvements in efficiency. There are other possibilities too. A rather similar context, where there are too many candidates for every possible pairing to be possible, occurs in some sports and games. Chess tournament organisers have developed a system for just these circumstances, called 'Swiss rules'[4]. Here the first round of pairings is purely random, but in the second round winners are only matched against other winners, and losers against other losers. Since the main concern is just to pick a final winner, while keeping the other participants contentedly occupied, in the third round those with two wins so far will only be matched against each other. In principle the system continues like this until only one is left who has won every game, so that just n rounds are needed where n is the smallest integer such that 2^n is greater or equal to the number of competitors. In practice, those with only one defeat near the end may be allowed back into the competition for the ultimate prize. Meanwhile, every participant has played several games against opponents of quite similar ability.

In a way this amounts to a very crude method of estimating the parameter values; indeed the ratio of number of 'wins' to number of 'defeats' is used to generate the starting values for some computer programs that estimate Thurstone paired comparison parameters. The advantage of the Swiss system for us might be that such simple analysis is more robust than early parameter estimation, particularly in that it does not need all of the data to be linked.

We usually can improve on the Swiss rules, because we do not need to begin with a round of random pairings. There is generally some information already available about each candidate. In the English school examinations system we ask teachers to forecast the grade they think each student will get. There is considerable discussion about the usefulness of these predictions: some teachers clearly use them, in different ways, to motivate students, others just seem to be very inaccurate or biased, but it is certain that the forecasts are accurate enough to make the first round of paired comparisons considerably more efficient than random pairing would achieve.

If no such independent information is available I suggest we ask the students themselves to tell us what result they are expecting. They will have no reason to over- or under-estimate their potential, and will know that their forecast will be

[4] I am grateful to Donald Laming for describing this possibility to me.

checked.This information must be better than nothing, and might turn out to be very useful.

Step 3

Reference scripts from previous years are included in some comparisons

Towards the end of the process we can send out some pairings that involve one of the 'live' scripts and one script from a previous version of the same examination. These act like 'anchor items' in item banking; their function is to align the temporary scale that arises from the analysis of live scripts with the scale used to interpret the previous version of the examination. The result is automatic test equating. In the English tradition any activity of this kind (they happen mostly in the context of reviews and appeals) would normally use scripts chosen to represent the boundaries between grades, since we would be asking judges to decide whether a script was above or below that boundary, but in the paired comparison context this is not necessary. Any scripts can be chosen so long as they give a reasonable spread of quality across the scale.

Thus award meetings become redundant. These are the meetings of senior examiners that take place at the end of the marking process in order to agree on the scores that will be required for each reported grade – A to E or A* to G. In two days (usually) they read many scripts and try to agree which ones do and which do not show convincing evidence of 'A quality', or 'C quality' or 'G quality'. The meetings are highly stressful affairs, as there is very little time to reach agreements, and the decisions will have serious consequences for candidates. There are many things that can go wrong, from incomplete marking and missing scripts to disagreements between the judgements of the examiners and recommendations of the statistical processes that are used to check them, and there is little time to resolve conflicts. Those who are not familiar with the English system and its procedures can find full details in the mandatory Code of Practice (QCA, 2004). Every examining board would dearly like to find a way to avoid such meetings, and the solution may be to replace the risky 'direct' judgements of script quality by paired comparison judgements fitted routinely into the basic process for evaluating the scripts.

Step 4

'Grey zone' scripts are defined

In recent years pressure on the system in England has reduced the capacity to check scripts that are close to a grade boundary and in danger of being mis-classified. In the system I am proposing the combination of parameter and standard error provides an automatic way of identifying these scripts, by asking a question like:

Does $val_{script} \pm 1$ standard error *cross a boundary?*

If so, then the script can declared 'grey'. This is conceptually similar to some current activities that define a grey zone of uncertainty around a decision point, but with one important difference. Greyness here is defined individually for each student's script.[5]

[5] This is in fact similar to a procedure something used in adaptive testing, if results are to be reported in grades.

Step 5

Extra data are gathered to resolve the 'grey zone' scripts

Now comes another meaning of 'optimisation'; the system can automatically send out for more judgement only those scripts that need more data, those which are too close to a boundary. Once a script is known to belong in a particular category there is no point gathering more data on it, and the effort can be re-directed to collecting information for those scripts where it is needed. This principle of resolving the greyness of each script provides a sort of stopping criterion which would allow the system of distributing pairs of scripts to judges to be run more or less fully automatically.

Ultimately, of course, there will need to be a rule for stopping. There will be some scripts which are, truly, very very close to a boundary, and we probably cannot justify gathering the enormous amount of judgemental evidence that might be needed in such a case. One possible strategy to deal with this would be to send that script out two or three times more with archive scripts that are already defined as being on the boundary.

Other forms of comparability

The system can be extended further. As described above, the scale is anchored by reference scripts taken from previous versions of the same examination. But it is also possible to add some comparisons with scripts from different examinations being taken at the same time. In England this means that the comparability of similar exams from different boards can be ensured by adding a few scripts from another board, or that the equivalence of different systems like GCSE and International GCSE can be checked in the same way. *Thus inter-board comparability studies become redundant*

We might even add a few scripts from a quite different subject, either to monitor the equivalence of the two subjects or to aid in setting a new standard when a new examination is created. QCA researchers are currently exploring the possibilities and difficulties of getting judges to make paired comparisons across related subjects. *Thus inter-subject comparability studies may be designed into the system too.*

With these possibilities we are really stretching the method, perhaps beyond its limits. At least it would be necessary to alert judges, and possibly necessary to train them, before sending such pairs to them.

Quality control

As described, standard errors of estimation can be used to ensure that each script is located on the standard scale with sufficient accuracy; this is quality control in a quantitative sense. There are some other ways in which the system can easily be monitored to see if the judgements being made are good enough in a qualitative sense. It may be that this facility for quality control will prove to be one of the most important advantages of the paired comparison system.

As well as an estimate of its value parameter and standard error the analysis will report, for each script, a misfit statistic. Essentially this reports the amount of inconsistency in the various judgements that have been made of that script. It can act as a flag identifying scripts that judges find difficult to judge, and might be used to divert these scripts to particular, perhaps senior, judges. Another

possibility, taken from the Rasch tradition, is to use the misfit statistic to inflate the standard error, using the equation:

Real S.E. = Model S.E. * Maximum [1.0, sqrt(mean-square misfit)].

This equation is attributed to Stenner, cited in Wright (1995). On average, the mean-square misfit statistic will, of course, be 1.0, and scripts that are judged more inconsistently will have higher misfit values, leading to an increase in their 'real' standard error. The result of inflating the standard error in an automated paired comparison system will simply be that the script in question will be more likely to be considered 'grey' and so be sent out for more judgements.

In exactly the same way misfit statistics can be used to monitor the consistency of the judges. This time it would be more appropriate to use the 'flag' alternative, alerting the managers to the problem. Of course, the system can immediately stop sending pairs of scripts to that examiner while the problem is investigated; there is no forward commitment of the kind our present postal system involves, with large packets of scripts sent out at a time. If it is decided that a judge must be removed the impact – extra work – will also be spread around the other judges far more evenly than we can manage at present. At any time judges who are not judged misfitting can choose to pull out or ask for more pairs, fitting the 'supply' of script pairs to 'demand' from examiners automatically.

Does comparative judgement work?

When we first suggest to current examiners that scripts should be judged holistically they often react with scepticism, or worse. It seems natural to evaluate essays, paintings or some other objects in this way, but it is not at all obvious that the method can work for scripts consisting of relatively discrete and small elements.

We have more than ten years of experience of applying paired comparison methodologies in a variety of assessment contexts, mostly concerned with the summative assessment of school achievement in the British examination systems. Most of these have been comparability studies, experiments designed to explore the equivalence of similar examinations. The list of subjects includes:

A Level or AS *(age 17/18)*
Geography, Mathematics, Chemistry, Biology, Accounting, Psychology, Sociology, English, History, Media Studies

Vocational Certificate of Education/A Level *(age 17/18)*
Health & Social Care, Business Studies

GCSE - International and UK *(age 16)*
French, Dutch, German, Afrikaans, English, Mathematics,

World Class Arena - Maths *(age 9-13)*

Cambridge Proficiency in English *– ESOL (age 17 to adult)*
Speaking - Oral interview

Key Stage 3 *(age 14)*
English

These range from obvious 'judging' exams like the assessment of foreign language speaking ability or A Level English, which is entirely assessed by essays, to equally obvious 'counting' exams like GCSE Maths or A Level Chemistry (eg, Bramley, Bell & Pollitt, 1998). Some, such as the foreign language exams, include multiple choice items and other dichotomous items; others, like Business Studies, include coursework projects. The World Class mathematics test involved small but complex problem solving tasks administered both on paper and on computer. The analysis of one study is reported in detail in Pollitt & Elliott (2003a).

Furthermore, most of these studies required the assessors to form a holistic view of whole sets of work. In a British A Level, for example, there are typically six separate papers or 'units', each being in effect a separate examination lasting usually 60 or 90 minutes, and these may vary quite considerably in the mode of assessment and the nature of the evidence of achievement that they provide. The foreign language examinations involved combinations of papers in reading, writing and listening.

In several of the studies the examiners began with grave doubts about the feasibility of making consistent holistic judgements about their examinations, but in every case they agreed to try, and in every case the results from nearly all examiners were satisfactory. After the experience, almost all of them accepted that the method could work.

How does comparative judgement work?

We are therefore satisfied that paired comparison judgement methodology *can* work in a surprisingly wide range of contexts. Sometimes we wonder *how* it works, and it seems unlikely that the process of holistic judgement will be the same in every case.

In Thurstone's original work on comparative judgement the aim was to construct scales from instantaneous evaluations, such as by asking people which of two tastes they prefer, or which of two pictures they would choose. When we first applied this technique to educational assessment, and to the considered judgement of professionals, we were not sure that it would give us consistent judgement data, meaning data that maintain stochastic transitivity (If A usually beats B, and B usually beats C, then A will mostly beat C). In the first of our studies (Pollitt & Murray, 1993) we added a second dimension to the study to explore the nature of the judgements being made. I had noticed that there are significant similarities and complementaries between Thurstone's scale construction technique based on the comparison – quantitatively – of two objects and George Kelly's therapeutic diagnostic technique in which patients compare –qualitatively – two (or three) people they know (Kelly, 1955). Kelly asked the patient to describe the similarities and differences between these people, and considered that their responses showed what thoughts were uppermost in their minds, the *personal constructs* they use to make sense of the social world. The evidence we gathered in the 1993 study showed clearly that untrained holistic judges were very selective in what features of performances they paid attention to, and that these selections changed in a consistent way at different levels of the scale. Since then several of the comparability studies have also contained a 'Kelly' component (Elliott & Greatorex, 2002), but much more serious study is probably needed if we are to

convince all those interested in examinations that it does make sense to use holistic judgement to evaluate scripts.

The real significance of our experience so far, though, is that the method manifestly works in many assessment contexts, in that it generates data that are consistent and that all of the researchers involved (from all the main English and Welsh examination boards) have found credible. If judgement works with current exams, which are designed for marking, how well would it work with exams designed for judgement?

Advantages of judgement

That paired comparison judgement *can* work does not mean that we *should* use it in every examination but it does mean, I think, that we should consider whether or not the advantages it brings would make it a better choice than the present marking approach.

Reliability

The concept of reliability is well known but in our context it is much more useful to consider the related idea of *precision*, or the 'reliability' of the estimation of a single script's quality.

With current exams, precision is largely a function of the number of questions, or test length (Lord, 1959); this is unfortunate given the practical difficulty of collecting more data from students. In contrast, with Thurstone scaling, precision is largely a function of the number of comparisons that are made *after* the examination is over for the student. In principle this might allow us to operate with shorter exams, so long as there is enough in the performance to be judged to represent fairly the principle aims of the course. We need to reconsider the current assumption that the real 'measurement' takes place in the examination hall, that markers are a necessary evil, trained to not-think, to mark like automata, and ideally would be replaced by machines. In the Thurstone view, measurement takes place in the judgement process every bit as much as in the answering process.

A similar difference can be seen with adaptive testing, where the current model requires us to set additional targeted questions to a student until our precision criterion is met. In adaptive Thurstone scaling we would instead send additional targeted comparisons to judges until the same precision criterion is met.

Validity

In an award meeting, as mentioned earlier, examiners have to choose a single score as the minimum required for a student to be awarded, for example, an A grade. Amongst the scripts they read will be several given exactly that score, or a little higher, which they have judged not worthy of A, and also some scoring less that are judged to be good enough. That is, although marks correlate with judged quality the correlation is far from perfect. It is clear that the total score does not capture 'quality' as well as we would like. In a paired comparison judgement system this problem and its consequent unfairness would vanish.

Currently, examinations are designed for 'reliable' marking. There is some evidence from our observations of examiners that questions are distorted by the need for reliable marking by single isolated markers. We are currently exploring the effect on validity of freeing examiners from the need to ensure reliability.

In principle I believe that direct judgement *should* be more valid than indirect scoring, since the construct that we are trying to assess becomes the actual criterion, or criteria, that examiners will use for their judgements.

The examiners who have been involved in our comparability studies certainly accept the validity of direct judgement as the basis for judging whole qualifications. It seems inevitable that this perception will be more true if comparative judgement is applied at the level of units rather than whole qualifications.

Quality control

The paired comparison system offers detailed quality control, of a kind not currently available, even when item level data are collected. We begin again with the principle that every judgement of a pair of scripts is independent. A report can be obtained that evaluates each decision:

> Judge 7 says: 9: "C4" beats 29: "W4" Calculated probability is 0.086 Standardised residual is 3.26

The report above flags a surprising decision by one particular judge, who was comparing two particular scripts. Of course some surprising judgements are expected, but these 'standardised residuals' are the raw data for all of the analyses of misfit described earlier, and of investigations of putative bias.

Bias control

The system makes it relatively easy to monitor routinely for various forms of bias. Any feature of a comparison that can be routinely coded into the data can be the basis for an analysis of bias. Examples of possible sources of student level bias that can be explored in this way include the quality of handwriting, the use of particular layout conventions, or a tendency to reproduce teachers' notes. At the level of examiners we can explore bias arising from methods of making a judgement, perceptions of 'new' versus 'old' (in comparisons of standards over time), of the time of day or night that examiners make their judgements, and of the influence of other scripts adjacent to the ones being judged.

So far, the only detailed study of this kind has been an investigation of the possible 'home/away' bias in comparability studies, where it was hypothesised that judges will be more favourably inclined towards scripts from their own board as opposed to those from other boards (Pollitt & Elliott, 2003a). The analysis showed that biases of this kind can always be detected so long as not all of the judges are equally biased in one direction.

Certification on demand

A final advantage I would like to suggest is that paired comparison judgements provide a rather easy way to offer certification of achievement whenever a single student or a small group of students request it. The principle of the independence of comparison judgement data means that any single script can be judged against any available other scripts. These may be others submitted at more or less the same time or a standard set of reference scripts, or any mixture of these two types. The result is that, for any single student, a 'certificate' can be issued as soon as a script is successfully calibrated. The

stopping criterion for precision, the resolution of 'greyness', is all that is needed to determine when a certificate can safely be awarded.

Concerns

There are some concerns that need to be resolved before a paired comparison judgement system can replace marking. The most important is probably whether or not sufficient precision can be achieved without excessive cost.

Costs

From our experience with comparative studies we estimate that we could make about **10** comparisons per script (this figure is based on the amount of time needed tomake judgements in A Level studies, compared to the time we pay examiners for marking similar scripts). In the comparability studies we generally use more than **50** comparisons per script. However, it is clear that we collect more data than are needed just to scale the scripts, and the rest of the data are needed only to carry out the fairly strict checks on quality that a politically sensitive study of this kind needs. For simple scaling we think that we would always be safe with about **25** comparisons per script.

'Intelligent' choice of pair partners, using the suggestions referred to above, would reduce the requirement further to, I think, less than **20** on average. It is also likely that moving to smaller units for comparison than the large and often heterogeneous sets that constitute a whole A Level examination would further lower the number required. Nevertheless, I think that some other source of data may be required to give enough precision for summative assessment.

Sources of additional data

If students are assessed electronically and submit their scripts as computer files then computer ratings of the scripts can be made, using one of several auto-marking programs. These ratings could be combined with judgements, suitably weighted, to give a more authoritative total estimate of value.

But a more interesting, and more general, approach would be to incorporate teacher's rankings of their students as part of the system. It is often supposed that teachers have a more complete knowledge of their students' worth, based on multiple informal and formal observations of their behaviour and performances; particularly in primary schools teachers' rankings have often been used as criterion variables for standardised tests of reading or arithmetic (See, for example, references to Bowman's Test of Reading Competence, France's Primary Reading Test, the Southgate and Dennis Young reading tests, the Basic Number test, the Leicester and Nottingham tests and France's 'Profile' in mathematics, all in Levy & Goldstein, 1984). In our current research it seems that secondary school teachers may often be equally confident in their ability to rank their students.

The utility of this for a paired comparison judgement system is considerable. A rank ordering by a teacher can be considered as a local Guttman scale; if a teacher ranks 20 students this can be interpreted as 19 extra comparisons for each student, since the one ranked top 'beats' the other 19, the one ranked second 'beat' everyone except the top one, and so on. Simply adding these implicit comparisons to the general data set will often give us enough data to ensure adequate precision. How we should calculate, or estimate, standard

errors in such a system, where the comparisons are no longer all independent of each other, is an unresolved problem, but the potential is certainly there to add teachers' evidence routinely to an otherwise 'objective' system of summative assessment. Many educationists would applaud such a move.

At this point it is worth mentioning another variant on the basic Thurstonian procedure. In one as yet unpublished study, aimed at test equating, we asked judges to rank order sets of ten scripts – five from a current test and five from a previous version – instead of making all the direct paired comparisons between them. Once again it seems that the judges found the task straightforward, the data were consistent, and the analysis was very convincing. In the next few weeks we will discover whether or not the outcome was completely acceptable. If it is, then we will have initial evidence that there are other possible ways of increasing the net amount of data in the system. In the meantime we are planning another formal test of this system as an alternative to the much unloved award meeting in the context of A Level this autumn.

Quality indicators

If a paired comparison judgement system is to replace marking we will need to develop statistics that adequately indicate the quality of particular Thurstone scales, statistics equivalent to (though, I hope, better than) indices of reliability. Since 'reliability' is so intimately bound up with the traditional test model, with its notion of a true score, something different is needed here, and I suggest that it will be something like a mean standard error of estimation at each grade boundary, perhaps expressed as a fraction of the width of a grade. Of course, a statistic like this will still be a function of several factors, including the number of judgements and their misfit or quality, and hence of the quality of the team of examiners as a whole, but it will at least indicate the degree of confidence that users can hold about the overall assessment procedure.

In a similar way we need to develop better quality statistics for judges. This is a practical issue, since we will want to identify, and remove, judges who deviate significantly from the average of the whole team. But this highlights the source of the problem; misfit statistics, as defined above, are normative. A reasonably good judge may misfit if the other judges are unusually consistent, while a rather poor judge may fit well enough if most of the others are somewhat inconsistent. We can start with normative statistics like the mean square, and get a system operating, but I think we will soon need a more stable and defensible basis for depriving some judges of work when we continue to employ others who may be no better.

Understanding judgement

The question was raised earlier of just how judges will make their judgements in different contexts. There are some more immediate concerns about how they will be *perceived* to make them. It is likely that in a system such as I have described there will quickly be some challenges and accusations.

Will the judges be unduly influenced by superficial features like handwriting? Will they vary in an unpredictable way in their response to particular features of style and content? Perhaps they will prove sensitive, or insensitive, to political demands that more, or less, credit should be given to spelling accuracy or elegant turns of phrase?

Perhaps more seriously: will they make 'hasty' judgements? In the current British systems examiners cannot simply look at the first one or two pages and ignore the rest, because checkers will quickly spot scripts that do not show marks for every question up to the last page. But in a paired comparison judgement system there will be no such direct evidence of 'due care'. In fact this *may* not be a problem, so long as most examiners behave properly; then, *in principle* at least, fit statistics and bias analyses should be able to control for these problems. We will need experience, though, to find out if we can collect enough data to identify judges who misbehave in this sort of way.

Training judges

My main concern in recent years (as reported, for example, in Pollitt & Ahmed, 2001, Ahmed & Pollitt, 2002) has been to improve the quality of the quality of the questions that we set students. If the questions are flawed the performances we collect will not be valid evidence of achievement. As we have said:

> *a question can only be valid if the students' minds are doing the things we want them to show us they can do.*

Currently, the requirement on examiners that they should write a mark scheme when they write the question exerts a valuable discipline, requiring them to try to anticipate the sorts of resonse their students will make to each question. There is a danger that the absence of marking will remove this discipline. It must, I think, be replaced by a sort of generic mark scheme in which examiners will express the general qualities they expect to see in a response with particular reference to the demands of each specific question.

It may then be necessary to train examiners quite explicitly in the use of generic schemes of this kind in order to ensure that they do follow the consensus rather than an idiosyncratic view of achievement in their subject domain. But there is a possible alternative. In this system the best judge is the one you never notice, the one who is never flagged by the system as a misfit. If we are prepared to trust this automatic way of monitoring judges then we may not need explicit training. We could instead allow any or every teacher to judge a few comparisons, monitor the results, and then keep the ones who are consistent with the consensus.

Suppose we do adopt that approach. We might implement it by making it compulsory - part of the package of entering students for the examination – or at least expected, that every teacher carries out a few comparisons of scripts (not, presumably, involving their own students). It would not be an intolerable burden to require ten comparisons from each teacher; indeed it is likely that most teachers would welcome the opportunity to see how a few students from other schools answered the questions.

This would (a) generate about 10% of the comparisons we need, (b) let us identify potential judges and screen out unsuitable ones, (c) improve the match between teaching and assessment (especially if we can design automated feedback), and (d) help to ensure public acceptance of the results.

Public trust

This leads to my final concern: will a system of this kind be accepted?

Currently, the public accept marks, though I can think of no good reason why they should except that they are used to them. Compared to the present system I would argue that it must be 'better' if each student's work is judged by ten or more judges instead of just one (or occasionally two); the risk of unfairness is surely reduced. It will be further reduced by the careful use of fit statistics to identify any area of concern in the data, and the collection of further data whenever a problem is identified.

In Britain, at least, considerable importance is attached to the idea of *transparency*, which is supposed to protect students against unscrupulous manipulation of results by assessment agencies in order, it is sometimes thought, to generate politically acceptable overall results. The processing of Thurstonian data is not immediately transparent, but we would be able, in any case where an appeal or complaint is likely to arise, to send the script in question for direct comparison against recognised reference scripts known to be located at grade boundaries. If we answer a challenge with evidence from direct comparison with multiple reference scripts, it is unlikely that the challenge will be sustained. It is well established that law courts do not challenge professional judgement so long as proper procedures have been followed.

Conclusions

We currently operate a summative assessment system in which judgement of the quality of the achievement of students is compromised by an unnecessary concern for the reliability of marking by a widely distributed team of markers whom it is difficult and expensive to monitor well. Yet the alternative of paired comparative judgement has been shown to work adequately with these same current exams. Future exams could be designed to meet the summative purpose directly – and changes of this kind would only make things better for a judgement system.

In addition, a paired comparison (or rank ordering) system would provide much more precise quality control over the essential processes of evaluating students' performances; this quality control can only improve the system further and increase public trust.

Marking was invented out of a fear that holistic judgement was unsafe in a time when there was little chance to control the quality of the judges. It grew with concerns for objectivity and transparency and an adherence to procedure rather than validity. The concept of reliability and concern about the apparent unreliability of judges made the system pay less attention to issues of validity and purpose. The development of theories that were wholly about items and item scores drove these trends to such an extent that wise human judgement found less and less place in examining. The more we came to rely on marking the more we distorted our summative instruments to make reliable marking possible, and the more we threatened the validity of the whole process.

Now we have an alternative. Thurstone's methods have waited 80 years, but scanning and IT transmission technology are at last making it feasible to put in place a system that will apply paired comparison judgement to school examinations.

In what could be seen as a return in spirit to the mediaeval tradition of summative assessment we will soon be in a position to make defensible and

accurate judgements of 'mastery' by asking our experts to judge what will be, quite literally, 'masterpieces'.

References

Ahmed, A & Pollitt, A (2002). *The Support Model for Interactive Assessment.* Paper given at the IAEA Conference, Hong Kong, September. Available at http://www.ucles-red.cam.ac.uk/conferencepapers.htm

Bramley, T, Bell, JF, & Pollitt, A. (1998) Assessing changes in standards over time using Thurstone Paired Comparisons. *Education Research and Perspectives*, 25, 2, 1-23.

Elliott, G & Greatorex, J. (2002) A fair comparison? The evolution of methods of comparability in national assessment, *Educational Studies*, 28, 3, 253-264.

Hambleton, RK & Swaminathan, H. (1985) *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Kelly, GA. (1955) *The psychology of personal constructs*, vols I and II. New York, Norton.

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

Lawley, DN. (1943) On problems connected with item selection and test construction. *Proceedings of the Royal Statistical Society of Edinburgh*, 61, 273-87.

Levy, P & Goldstein, H. (1984) *Tests in Education*. London, Academic Press.

Linn, RL (Ed.) (1993). *Educational Measurement*. Third Edition, New York: Macmillan Publishing Co.

Lord, FM & Novick, MR. (1968) *Statistical theories of mental test scores*. Reading, MA, Addison-Wesley.

Lord, FM. (1959) Tests of the same length do have the same standard error of measurement. *Educational and Psychological Measurement*, 19, 233-39.

Pollitt, A & Ahmed, A. (2001) *Science or Reading?: How students think when answering TIMSS questions*. Paper given at the IAEA Conference, Rio de Janeiro, May. Available at http://www.ucles-red.cam.ac.uk/conferencepapers.htm

Pollitt, A & Elliott, G. (2003a) *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.ucles-red.cam.ac.uk/conferencepapers.htm

Pollitt, A & Elliott, G. (2003b) *Finding a proper role for human judgement in the examination system.* Paper given at the QCA 'Comparability and Standards' seminar, Newport Pagnell, 4th April. Available at http://www.ucles-red.cam.ac.uk/conferencepapers.htm

Pollitt, A, & Murray, NJ (1993) *What raters really pay attention to*. Language Testing Research Colloquium, Cambridge.

QCA. (2004) *GCSE, GCSE in vocational subjects,GCE, VCE, GNVQ and AEA: Code of practice 2004/5*. http://www.qca.org.uk/about/board/6295.html

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

Wood, R. (1991) *Assessment and testing - a survey of research*. Cambridge: Cambridge University Press.

Wright BD (1995) Which standard error? *Rasch Measurement Transactions*, 9(2), p.436-7. http://209.238.26.90/rmt/rmt92n.htm

Wright, BD & Stone, MH. (1980) *Best Test Design.* Chicago, MESA Press.