Finding a proper role for human judgement in the examination system.

Alastair Pollitt
Gill Elliott

Research and Evaluation Division
University of Cambridge Local Examinations Syndicate
1 Hills Road
Cambridge

4 April 2003

In this paper we extend the discussion in the background paper, to consider some potential applications of the Thurstone method of Paired Comparisons. Because the analysis is modeled at the level of *individuals* rather than of *groups* it is much more powerful than the old method. There are two parts to it which may be called **The Model** and **The Misfit**. The model part constructs a single scale showing the relative value of all the scripts included, while the misfit part allows investigation of any effect that may perturb that common scale.
Together they constitute a general system within which comparative data can be used to construct and validate any set of assessments.

Contact details
Alastair Pollitt & Gill Elliott, RED, UCLES, 1 Hills Road, Cambridge, CB1 2EU.
pollitt.a@ucles.org.uk and elliott.g@ucles.org.uk.

1 Summary of background

Before considering the role of human judgement in both monitoring and maintaining standards in the future, it is worth revisiting the past to remind ourselves of the theoretical basis upon which both current and previous practice developed – reasoned consideration of the past can significantly determine the evolution of practice for the future. Two principle methods of using human judgement to investigate comparability between examinations have been used in the last quarter century. Both have considerable practical similarities – using balanced teams of examiners from each examination being compared to judge candidate performance at key reference points (usually borderlines), and generally carried out over the course of a two or three day residential meeting. There are, however, significant theoretical differences between the two models.

Old method – Home & Away Ratification

Awarding bodies began to use a structured method for eliciting examiner judgements about the relative comparability of parallel syllabuses in the mid-1970's, and the methodology which developed continued to be used for the next twenty years. The methodology was based upon two premises:

i That human judgement represented the best means of 'carrying' a standard from one examination to another, because only a human mind can hold onto the essence of the existing standard whilst at the same time adjusting their opinion according to the differing context of the second performance being judged. Examiners (particularly senior examiners) have an inherent knowledge of the standard of the quality of work usually seen in their syllabus at given points of reference – i.e. borderlines. This remains a central factor in comparability work of this kind to this day.

ii That each judge would tend to be loyal, even biased, toward their 'home' examination – thus if all else proved equal they would judge their own examination as more stringent than an unfamiliar exam.

Examiner judges were asked to concentrate upon the standard they would expect in their home examination at the given reference point, and then to judge whether each candidate's work was on the borderline, above or below it. Each judge carried out this task on home scripts first, to further entrench the home standard.

Results were compiled in much the same way as sports results with tables of wins and losses converted to scores. If both teams of judges detected an equal home advantage, a draw could be declared.

New method – Thurstone Comparative Judgement

In recent years the methodology has changed: the same teams of examiners are recruited but make their judgements in an altogether different way. Instead of judging the scripts against their internal standard, they compare one script directly

with another, and are required to determine which of the two scripts shows evidence of higher value. A Rasch analysis can then be carried out which combines all the judgements (by estimating a parameter value for every script in the study) and orders the scripts onto a single scale. This approach exploits a methodology called *comparative judgement* proposed by L. L. Thurstone (1927) for constructing scales to measure psychological phenomena.

Power

A great strength of the Thurstone paired comparison approach is that the data provide the means to test both **model** and **misfit**. The direct comparisons produce the model (the scale and the rank ordering of the scripts), whilst the differences between these judgements and the model's predictions produce a measure of misfit for each comparison. Misfit statistics can be used in a huge variety of ways, to investigate any systematic effects which may be suspected, whether these apply to scripts or to judges; this is something we will return to later.

A single trait

Scale construction

Thurstone designed this methodology for psychologists to create, *de novo*, a scale for measuring the perceived value of any objects, concrete or abstract, on scales which might measure 'attractiveness', or 'goodness', or 'radicalness'. It was used, for example, to explore people's preferences for different kinds of food, or to construct ethical scales, or to explore perceptions of politicians and their policies. Whatever the objects used, participants in the study implicitly constructed a single trait to underlie them all, and located each pair of objects relatively on that scale.

In his experiments participants made instant judgements about the objects, and it is not immediately obvious that the same methodology can be safely applied to the extended and deliberate considerations involved in comparing examination scripts. Experience so far, however, suggests that it works surprisingly well: well enough to encourage us to explore further possibilities.

Minority syllabus

When we deliberately make the assumption of a common underlying trait more difficult, as happens when we include scripts from two or more syllabuses, there is a possibility of a simple bias against a syllabus that is somewhat different from the others. Suppose that we include four syllabuses, of which one differs by emphasising skills or knowledge not given as much value by the others. It is likely that scripts from that syllabus will be valued less by the majority of the judges, who are used to the values of the other syllabuses; the result will be that that syllabus appears to set a lower standard.

Model and Misfit

In such a case, however, where the *model* seems to fail to be fair, the second part of the system, the *misfit* shows us what is happening. All of the judges from the minority syllabus will show a bias in favour of scripts from their own syllabus, because they – honestly – value those scripts relatively higher than the majority do. This analysis clarifies the real issue, and should lead to a more informed discussion of the relative validity of the different forms of assessment. We can conclude that a 'Home & Away' analysis should always be included in any comparability study.

System

The example above shows the importance of the two parts of the system. Any comparison across two examinations, and even within one where other factors may play significant roles, should consider both parts. First a quantitative analysis will show how a (reasonably representative) set of experienced examiners (the judges) actually value the awards made by the different assessments. Then misfit analysis will explore the validity of the conclusions of the quantitative study and elucidate the effects of any hypothesised disturbing effects.

Kelly's Personal Construct analysis

It is worth pointing out the remarkable congruence between Thurstone's method and George Kelly's *personal construct* approach to psychology (Kelly, 1955).

Kelly viewed a *man* as a *scientist* (these were days before political correctness), each one exploring the world for himself and coming to understand it through a process of hypotheses formation and testing; successful hypotheses become the basis for action and, consequently, the basis for conceptualising the world:

> "*A person's processes are psychologically channelized by the ways in which he anticipates events*".

Kelly saw this as particularly important in the realm of inter-personal relationships, and viewed personality as a set of constructs and working hypotheses formed through experience. The essential process of constructing the view of people and the world was one of comparison; for him it was the differences we see between one person and another that lead us to predict how each one will behave.

This led him to suggest an approach to psychological research which is widely favoured today. The researcher asks a subject to describe three people they know well, first describing how one of them is different from the other two, who are more the same in this respect, then repeating the process with each of the others singled out for comparison with the other two. The theory is that the features the subject uses spontaneously to distinguish one person from others will be the ones that are most salient for the subject, the ones he uses to construe the social world.

In his writings Kelly notes that it is not strictly necessary to use three 'stimuli'; comparing just two is enough to elicit salient constructs since there is an implicit normative presence of the rest of mankind at all times. Thus we come to a useful complementarity of two methodologies: both Thurstone's and Kelly's methods

depend on the comparison of two 'objects' to elicit something meaningful to the judge about the differences between them, in the first case a quantitative judgement of which one has 'more' of a certain quality such as 'ability at history', in the latter a qualitative but prioritised description of the ways in which they differ. The combination of the two can lead to an illuminating analysis of the nature of the traits that judges actually use, just what they 'see' in examination scripts. (Pollitt & Murray, 1995)

Qualitative Differences

For these reasons it has become common to incorporate an element of Kelly's approach to comparability studies, usually using a form of what is called 'repertory grid analysis', a rather more standardised research technique developed by Kellyian psychologists (for more information on Kelly and grid analysis see www.repgrid.com/pcp/; or www.enquirewithin.co.nz/HINTS/skills2.htm).

2 Applications 1 Bias analyses

In this section we comment briefly on how the *misfit* part of the Thurstone paired comparison system can be used to explore perturbation in the data collected. It is possible simply to 'trawl' misfit data looking for patterns, and this may sometimes lead to valuable insights. For example, Abdullah (1989) showed how factor analysis of questionnaire data may be improved by removing the dominant first factor and analysing the residuals instead of the raw data, and the same may be true here. Even so, it is generally considered good practice to formulate hypotheses for study before analysing the data, and here we identify some hypotheses that might be explored in this way.

In all of these it should be remembered that *bias* is a technical term in statistics referring to any systematic source of error or misfit in an analysis. There is no necessary implication of moral or ethical fault implied here in describing a pattern in the data as evidence of bias.

A Bias – Home & Away

The background paper showed how bias hypothesised to arise from the examiner's experience of one particular syllabus or approach may be evaluated. In discussion below we describe why this may be an essential part of any comparability study.

B Bias – Ascriptive variables

'Home-ness' is just one of many variables that can be used to describe candidates, their scripts or the judges. Others that could be investigated in a similar way include the sex, ethnicity or age of either candidate or judge, the type of school the candidate comes from (in whatever descriptive systems we choose to use) and – importantly – any combination of these. Since the residual data consist of one quantity for each judgement the analysis simply involves accumulating all of the residuals for whatever subset of the data we are interested in, and evaluating it for

significance. The subset we choose to be interested in may be defined in any way that amounts to stating a hypothesis.

C Bias – Handwriting

One variable worth singling out is handwriting. Previous attempts to investigate the impact of handwriting on assessment have always been carried out at group level, in effect using the hypothesis that every judge is biased against poorly written scripts to a small degree. Paired comparison analysis is an individual modeling analysis, and the hypothesis is effectively tested independently for each judge. In fact, this method will fail to detect bias if it is true that every judge is equally biased, since in this case 'handwriting' simply becomes one of the demands that together constitute the trait being assessed: the validity or otherwise of this demand is of course contentious, but it is arguable that there is no unfairness involved since the criteria are the same for everyone. But if some judges are more affected by poor handwriting than others, then there is clear unfairness since the rating a script gets will vary across different judges; this is the sort of bias that the paired comparison method will detect and describe.

D Judge Adaptability

One interesting hypothesis is that judges will change their behaviour as they progress through their series of judgements. We have some indications of this, in that judges occasionally ask if they can go back and change earlier decisions. There are several ways of testing for trends over time in the residuals, depending on exactly what concern is being checked.

E Judge Training

One direct application of this would be in the training of judges. We would expect judges who are new to the task to misfit more than experienced ones, in general, as they adjust their 'home' view of the trait being assessed to take account of the views of other syllabuses. If this is the first experience they have of other syllabuses we might expect a clear effect, though increasingly we are finding that examiners who take part in comparability studies have already experience of more syllabuses than the one they are supposed to represent.

In a training context we might choose to present particular scripts that illustrate particular problems for examiners, and provide feedback related to the judgements the trainees make of them.

3 Applications 2 Scale construction

The applications above all assume that there are two or more identifiable groups of scripts and/or judges, and arise out of the original intention to seek a better way to carry out studies of comparability. If, on the other hand, we remember that Thurstone's intention was to construct scales to measure a single trait then some other, more radical, ideas suggest themselves.

A Construct this year's scale for an examination

- Replace marking

The purpose of the marking process – in the present scheme – is merely to create a rank order of the candidates. Whatever questions each one chose, if choice exists, and whichever ones they got right or wrong, the end result is a total score, a number which locates the student at some point in the relative ordering of all candidates in terms of 'how much' they know of the subject. The score is not in any significant sense a measure; a score of 50% does not mean you 'know' half the subject or twice as much as someone who scores 25%. Nor is someone who scores 75% better than you by the same amount as you are better than the one who scored 25%. The raw score scale is merely a rank order: in statistical terms it is an *ordinal* scale that approaches but does not achieve *equal interval* status. Marking is supposed to generate a rank order with as much *reliability* as possible.

Thurstone's methods (paired comparison was only one of about 12 he described) all do better than this, creating scales with genuine equal interval properties. But, more importantly, the scale generates its rank order with as much *validity* as possible. Every judgement is made wholly in terms of the relative 'validity' of the scripts rather than through a proxy 'score' which we all know is an imperfect representation of 'value'.

If all scripts were scanned optically the images could be sent electronically to markers in any desired combinations for comparison on-screen. The technology exists to create such a system, pilots have been carried out, and we may have a fully electronic on-screen marking system in just a few years (for the moment the costs are still rather too high). We estimate that, in a Thurstone system using on-screen comparisons in place of marking, each script could be judged about 7 times for the same expense as at present. Whether this is enough to achieve the accuracy we need will be considered further below, but it is at least clear that it is possible in principle to replace the whole marking process with paired comparisons.

B Include archive scripts

- Replace awarding

Why stop with marking? The earlier discussion showed that comparability can be checked by comparing archive scripts with current ones. If paired comparison were the method for setting up this year's rank order then there would be no need to wait for *post hoc* comparability studies: the 'ranking' process could be seeded with archive scripts in such a way that boundaries would be automatically determined at the same time. The savings in time as well as cost would be considerable.

It's not clear yet whether all, or only more senior (?), examiners should be involved in standard setting judgements, nor how the critical judgement pairings should be selected but, again, it is clear that this is possible in principle.

C Include other syllabus scripts

- Replace comparability studies

Electronic script marking has many advantages, and electronic paired comparison would share them all. Particular scripts can be fed automatically to particular examiners at any time, as part of their routine processing of scripts. In order to carry out comparability studies it would be necessary only to warn and prepare some examiners for the task before setting them cross-syllabus pairs to judge.

At present examiners prepare by studying the syllabus documentation, question papers and mark schemes associated with the 'away' examinations in order to understand the demands the students face in gaining marks on the questions. If we have abolished marking, however, there will be no mark scheme, and preparation will need to be slightly different.

It is generally considered good practice to write a question and its mark scheme simultaneously. Our research on question writing shows that an essential part of writing a good question is the process of anticipating the responses that candidates will make: what we call predicting the outcome space. Whether we need marking schemes or not we will continue to need this description of the expected outcome space as part of the writing process and as the material for training the examiners we are proposing in place of markers and comparability judges.

Would a board be prepared to use judgemental data from such a comparability exercise as part of its own awarding process? If they are, then comparability studies would simply become an operational process of diverting a few scripts from this (home) examiner to that (away) one some of the time, and receiving a few others diverted in exchange. The comparability process would probably use more scripts than at present but make fewer judgements on each.

D Include other subject scripts

- Equivalencing

Again – in principle – this idea can be extended to include the comparison of scripts from examinations in other more or less related subjects, to check on the relative standards of exams in these different areas. But we are really stretching Thurstone's original idea to the limit now; the further we go from constructing a single homogeneous scale the more preparation there would need to be, the slower and more difficult it would be to make the comparisons, and the less consistency we would expect to find. And probably the less people would believe it.

E Include other modes ?

The last of the comparability problems we listed in the background paper concerned the comparability of assessing the 'same' subject in different modes: the 'Alternative to Practical' paper compared to current practical exams, or the use of computer simulated science experiments in place of paper-based questions, or indeed any case where 'paper and pencil' are replaced by 'on screen' tests. It seems

unlikely that any human judge will be able to balance fairly the many changes in skills and in the level of demands in these skills that are involved.

F Include prelim scripts

- Replace forecast grades

One further possibility would be to include other evidence of achievement to check the consistency of the ranking process. If, for example, we were to include scripts from students' school based prelim exams we would have an automatic check that students are not performing unusually well or badly in the examination, the role that teachers' forecasts of students' grades are supposed to fulfil but rather fail to do, in England at least. But see below for a better idea.

4 A new examining system?

In this final section we imagine how these proposals might be brought together to create a new system for examining, one which might overcome some of the problems we face every year with the present system. Many of the details need considerable further work, but everything we describe could be in place at least in principle within just a few years.

Teachers' rank orders

The starting point of our system is that teachers should play a greater role in determining the outcome of their pupils' school careers. In several European countries, most notably Sweden and Germany, teachers are trusted to award results with little or no external interference, but we are proposing a more controlled role. It is widely accepted that teachers are able to rank order their pupils with more validity than any external examination, and we suggest that boards require centres to submit rank orderings of their pupils, as close as possible to the exam date, drawing lines where they expect the grade boundaries to divide the pupils. This is, of course, traditional in Scotland.

Scale construction in place of marking

Next we would replace marking with Thurstone paired comparison judgements, using on-screen marking of scanned scripts to optimise the design, setting up comparisons of matched scripts as in current comparability studies. The initial pairings could be set up using the teachers' forecasts, but these would be revised in the light of early comparisons.

Moderation of rank orders

At this point two options arise. In one approach, analysis of the teachers' rankings as *ordered sets* could be used to merge them into a single overall rank order. This is based on the idea that the rank ordering within a single class can be converted into a set of paired comparisons, with the pupil ranked first being awarded a 'win' over

each other pupil, the second a 'loss' to the first and a 'win' over all the others, and so on. These 'results' are combined with the Thurstonian data for the full analysis.

Alternatively, the main analysis could compute mean rankings for each school's list and these could be used to moderate the teachers' rankings in a more traditional way.

+ scale equating

As described earlier, archive scripts would be introduced judiciously to ensure that standards were maintained from an anchor year; we suggest that a single set of anchor scripts be used for four or five years to reduce the risk of drift.

Other scripts could be introduced as necessary to ensure automatic comparability, not only between times, but also

- between subject areas,
- between boards,
- between syllabuses, and
- between units.

Review

The only reason ever to review a pupil's results would be a significant discrepancy between the teacher's ranking or forecast grade, and the result of the main paired comparison analysis. Any such discrepancy would in the first place be investigated by submitting the script for extra judgements.

The current review system is biased, in that the reviewed script is identified, and will only be revised upwards. In our new system judges would merely see another pair of scripts without knowing which, or even that either, of them was the subject of some kind of review. The cost of raising one script – the automatic lowering of another - would be obvious in every decision made, and misleading notions like "the benefit of the doubt" would simply never arise.

Only if a discrepancy was confirmed would extra intervention be considered. Rules would need to be set for how to handle such a conflict between forecast and examination. In England & Wales at present it is clear that examination performance always wins, except in very extreme special cases; in Scotland it is accepted that the teacher's judgement may sometimes over-ride the examination performance; with this new system we have the opportunity to consider a new balance between the two sorts of evidence of achievement.

Conclusion

If each script takes part in about seven comparisons, the results of the main analysis would be sufficiently accurate to support a system based fundamentally on the teachers' ability to rank their pupils appropriately. Indeed, if we accept this as the essential strategy, then we do not need to scan every script, since twenty-eight judgements on a randomly selected quarter of each teacher's set would be just as accurate in most circumstances, and more efficient.

Summative assessment should be about overall judgements, not about details. With this system marking of pupils' answers, with its emphasis on where and why each pupil did or did not get credit, will happen only in the context of formative assessment, where the focus is on how to improve a pupil's performance. Summative assessment will only look at the overall level of performance: marking should have no place in it.

The principal benefit of this new system would lie in the simple fact that it is, from start to finish, based on considerations of validity rather than ever relying on reliability as a proxy for validity as we do at present. Teachers would feel that their superior knowledge of the pupils' *true ability* was properly recognised; judges would be free to give the decisions they believe are *fair*, rather than being required to follow a marking scheme literally; reviewers would seek *justice* for a pupil rather than mere accuracy in marking a script.

A higher priority for validity is surely worth seeking.

References

Abdullah, W M R. (1989) *The effects of teacher attitudes toward students on teacher planning, instructional support, and teacher's effort in maintaining order in the classroom*. Thesis (Ph. D.) University of Chicago, Dept. of Education.

Kelly, G.A. (1955). *The Psychology of Personal Constructs*. New York: Norton.

Pollitt, A, & Murray, NJ (1995) What raters **really** pay attention to. In M Milanovic and N Saville (Eds) *Performance testing, cognition and assessment. Studies in Language Testing, **3***. Cambridge: Cambridge University Press.

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