CAMBRIDGE ASSESSMENT

Maintaining performance standards: aligning raw score scales on different tests via a latent trait created by rank-ordering examinees' work

Tom Bramley & Beth Black

Paper presented at the Third International Rasch Measurement conference, University of Western Australia, Perth, January 2008.

Research Division
Cambridge Assessment
1, Regent Street
Cambridge
U.K.
CB2 1GG

Bramley.t@cambridgeassessment.org.uk
Black.b@cambridgeassessment.org.uk

www.cambridgeassessment.org.uk

Abstract

Different versions of tests in the same subject at the same level, taken at different times, but where there are no common items or examinees, need to have cut-scores set at the same performance standard. The rank-ordering method is a new technique for equating the raw score scales on two tests by using expert judgment, whereby the cut-score on one year's test can be mapped to the equivalent score on later versions, via a latent trait scale constructed using Rasch analysis. The judgmental task involves experts making relative, holistic judgments about examples of student work from two tests and combining these into a single rank-order.

The research activities have consistently shown that the rank-ordering technique has produced outcomes reasonably similar to other standard-maintaining activities used in the UK assessment context (which use a combination of statistical evidence and expert judgment). The method is also flexible in the experimental designs it allows and applicable in a variety of standard-maintaining contexts, having been used with success to: i) Equate pairs of tests comprising short-answer items only, essay style extended items only, and a mixture of short and longer items; ii) Equate tests of differential demand (vertical equating by expert judgment); iii) Investigate, post-hoc, whether standards have been maintained over time, using 6 consecutive sessions (i.e. different versions) of UK tests.

This paper explains the theoretical background (Thurstone paired comparisons), summarises research conducted to date, and explores the theoretical and practical issues arising in using this rank-ordering method for maintaining performance standards.

Introduction

A task frequently faced by assessment agencies is that of setting cut-scores on the raw mark scale of a test. The cut-score represents the boundary between two meaningful categories (in terms of interpreting the outcomes) such as 'pass/fail', or 'distinction/merit', or 'grade A / grade B'. If the test is completely new then the cut-score setting task is (usually) one of 'standard setting' and one of a plethora of standard setting methods might be adopted (see Cizek, 2001 and Hambleton & Pitoniak, 2007 for reviews). It is recognised that the standard setting task involves values and professional judgment, and the different standard setting methods try to incorporate these in as valid and defensible a way as possible.

If, on the other hand, the test is one of a series of versions or forms, constructed from the same test specification, and the cut-scores on one or more of the forms have already been set, then the cut-score setting task can be considered to be 'standard maintaining', where the aim is to set the cut-score at a point on the mark scale that is in some sense equivalent to the other tests. For this, the methods of test equating and linking can be used (see, for example, Kolen & Brennan, 2004; Holland & Dorans, 2007) provided that data corresponding to an appropriate equating design is available, and that the assumptions of the equating method can be met.

One way to conceptualise the standard-maintaining scenario is via latent trait theory. The performance standard can be considered to be a point on the latent trait (Bramley, 2005a). Examinees at the same trait level will be expected to obtain different raw scores on two different tests unless the distribution of item difficulties on the two tests is the same. Thus if the latent trait value corresponding to the cut-score on the raw score scale of one test is known, the standard maintaining task is to find the cut-score on the raw score scale of the other test corresponding to this same value. This is illustrated in Figure 1 below.

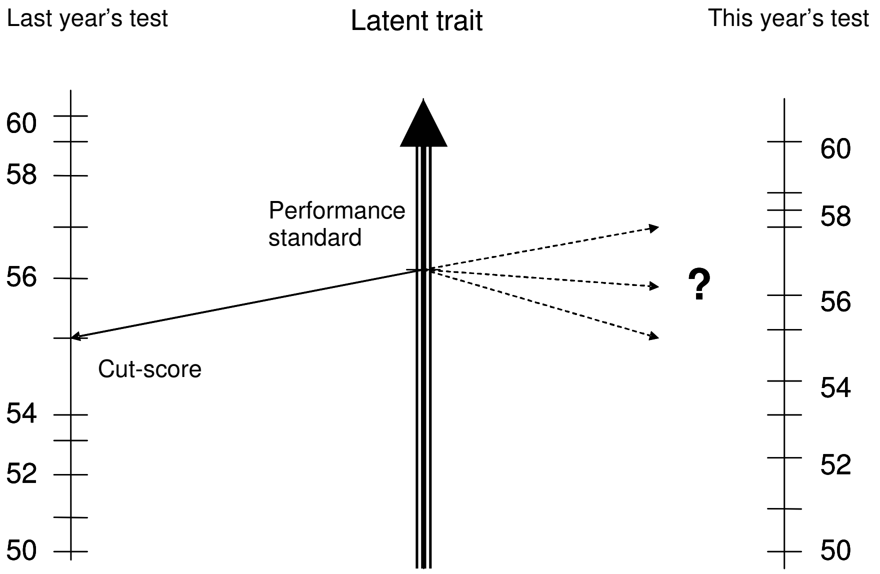

Figure 1: Illustration of the standard-maintaining task. The cut-score on last year's test was set at 55 marks, and the mark on this year's test corresponding to the same performance standard on the latent trait must be found.

This conceptualisation of the task underlies the Rasch approach (and other IRT-based approaches) to test equating. In order to estimate the latent trait values within the same frame of reference (i.e. on the same scale) it is necessary to be able to link the two tests together by a common element – either common examinees or common items.[1] The powerful conceptual framework provided by the Rasch approach created new possibilities for scale construction, item banking and adaptive testing (see, for example, Wright, 1977; 1984) and arguably still provides the best practical reason for using Rasch measurement.

However, in some assessment scenarios, it is not possible to equate tests in this way. This can be for a number of reasons:

- The tests are not considered suitable for application of the Rasch (or other IRT) models. This might be the case if there is a wide variety of item tariffs (e.g. a mixture of dichotomous items, 3-category polytomous items, 10-category polytomous items etc.), or a lack of unidimensionality (e.g. if subsets of items are testing different content or require different skills), or a lack of local independence (e.g. if more than one item refers to the same context, or if success on one item is necessary for success on another).
- The sample sizes are too small to obtain satisfactory parameter estimates.
- It is not possible to obtain a common element link. This can occur when the tests are very high-stakes, and item security is paramount (preventing any pre-testing), or when tests are published and the items used by teachers to prepare subsequent cohorts of examinees for the test (invalidating the assumption that common items would maintain their calibration across testing situations).

The first and third of these points very often apply in standard maintaining situations for high-stakes tests and examinations in the UK. The rank-ordering method (Bramley, 2005b) was developed in an attempt to use a latent trait framework for comparing performance standards when no common item or common examinee linking is possible. The method relies on expert judgment of examples of examinees' work (henceforth 'scripts'), giving it some similarity with examinee-based standard setting methods.

[1] In principle, it would be possible to link the parameter estimates without a common element link if extra assumptions were made – for example that the ability distributions of examinees taking each test were the same. However, this is not 'in the spirit' of Rasch measurement where the purpose is to achieve sample-free calibration that does not require such assumptions.

Thurstone paired comparisons[2]

The theory underlying the rank-ordering method is Thurstone's theory of comparative judgment (Thurstone, 1927a, 1927b), which is summarised below. When a judgment is made about an object (in this case a script) in terms of a specified attribute (in this case the quality of attainment apparent in the script) the object evokes a 'discriminal process' in the mind[3] of the judge. The same stimulus object does not evoke the same discriminal process each time it is presented, but rather these discriminal processes form a frequency distribution, which Thurstone defined to be Normal in shape, and thus able to be characterised by two parameters: the 'modal discriminal process', and the 'discriminal dispersion', corresponding to the mean and standard deviation of this distribution. Thurstone's insight was that a psychological scale measuring the attribute could be constructed from the distributions of discriminal process evoked by different objects. The discriminal mode and dispersion corresponding to a single object are inaccessible to observation. They can only be estimated when two objects are compared. Thurstone assumed that when two objects A and B are compared with respect to a specified attribute, the object evoking the discriminal process further along the psychological continuum would be judged as possessing more of the attribute. The situation is shown in Figure 2 below.

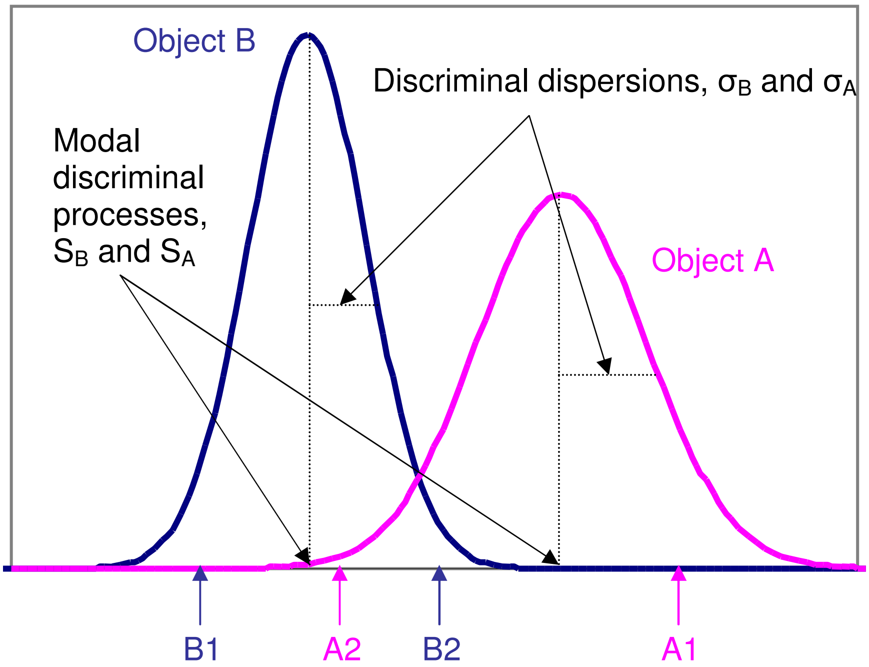
Object B
Discriminal dispersions, σ_B and σ_A
Modal discriminal processes, S_B and S_A
Object A
B1
A2
B2
A1

Figure 2: Example distributions of discriminal processes for objects A and B.

The outcome of the paired comparison judgment is therefore related to the distribution of the *difference* between the two distributions of discriminal processes for object A and object B. If this difference is positive, for instance when the objects evoke the discriminal processes at A1 and B1, we have the judgment 'A beats B', but if it is negative, for instance when the objects evoke the discriminal processes at A2 and B2, we have the judgment 'B beats A'.

[2] Parts of this section and the next section are adapted from Bramley (2008, in press).

[3] Thurstone believed that this psychological process would have a neurological correlate, but his theory was intended to operate at the psychological level, not the level of brain processes.

The mean of the distribution of the difference between the two distributions is the distance between the two mean discriminal processes – i.e. S_A-S_B, the scale separation between A and B. The standard deviation of this difference distribution, σ_{AB}, is given by the formula:

$$\sigma_{AB} = \sqrt{\sigma_A^2 + \sigma_B^2 - 2 \cdot r_{AB} \cdot \sigma_A \cdot \sigma_B} \qquad (1)$$

where:
σ_A is the discriminal dispersion for essay A,
σ_B is the discriminal dispersion for essay B,
r_{AB} is the correlation between the discriminal processes.

The final form of Thurstone's Law of Comparative Judgment is:

$$X_{AB} = \frac{S_A - S_B}{\sigma_{AB}} \qquad (2)$$

where X_{AB} is the deviate of the normal distribution corresponding to the proportion of judgments 'A beats B', and S_A, S_B and σ_{AB} are as defined above.
In words, the scale separation between two objects on the psychological continuum is measured in units of the standard deviation of the difference between the distributions of their discriminal processes.

Thurstone gave 5 cases of his law, each making more simplifying assumptions about the form of the denominator in equation (1) above. The simplest form (Case 5) sets r_{AB} to zero, and the discriminal dispersions σ_A and σ_B to be equal, making σ_{AB} a constant which can set as the (arbitrary) unit of measurement, giving:

$$X_{AB} = S_A - S_B \qquad (3)$$

In words: the scale separation between two objects is equal to the unit normal deviate corresponding to the proportion of judgments 'A better than B'. (Note that if this proportion is less than 0.5 the separation will be negative – i.e. B will be higher up the scale than A, as we would expect).

If the mathematically more tractable logistic distribution is used instead of the Normal distribution then equation (2) can be expressed as:

$$p(A > B) = \frac{\exp[a(S_A - S_B)]}{1 + \exp[a(S_A - S_B)]} \qquad (4)$$

where a is a scaling parameter which can arbitrarily be set to 1, (just as σ_{AB} is set to 1 in Case 5 of Thurstone's law of comparative judgment).

Logistic models are widely used both in general categorical data analysis where equation (4) is known as the Bradley-Terry model (Bradley & Terry, 1952; Agresti, 1990); and specifically in Item Response Theory in the field of educational measurement, where equation (4) has the same form as that of Rasch's (1960) model for dichotomous items. The connections between the Rasch model and Thurstone's Case 5 have been explained in detail by Andrich (1978).

In summary, the paired comparison method produces data which, when analysed according to the law of comparative judgment (Case 5), yields a value for each object on an equal-interval scale with an arbitrary origin and unit. The scale is equal-interval in the sense that the same distance between pairs of objects at different parts of the psychological continuum reflects the same probability of one ‘beating’ the other in a paired comparison.

The rank-ordering method

In a rank-ordering exercise, instead of comparing pairs of scripts, the judges are asked to put sets of more than two scripts into a rank order.

The procedural details of a rank-ordering exercise have been described in depth in Bramley (2005b). The following summary describes how it has been applied in work carried out to date (Bramley, 2005b; Black & Bramley, in press; Gill et al., submitted) .

Scripts are selected from the two (or more) tests to be compared such that the whole effective mark range is covered. Script mark totals and individual question mark totals (if feasible) are removed from the scripts which are then copied as many times as necessary for the study. Packs of scripts are compiled containing scripts from both tests. In most of the studies carried out to date, packs of ten scripts have been used containing five scripts from each test. The scripts within each pack can vary in both the range of marks covered, and in the degree to which the mark ranges from each test overlap or are offset. Each pack contains a unique selection of scripts, but there are many common scripts between the packs allowing the entire set of scripts to be ‘linked’, as illustrated in Figure 3 below.

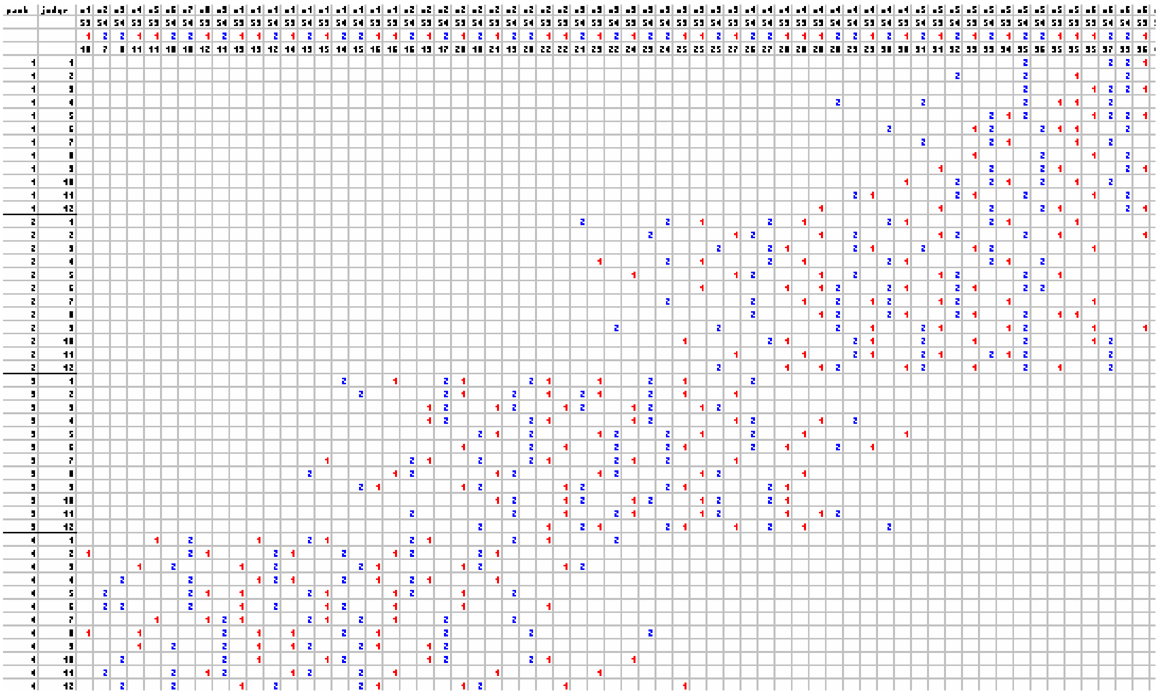

Figure 3: Example of part of a script allocation design in a rank-ordering study.

Each judge is given a number of packs covering the whole mark range. Some effort is made to minimise the number of times each judge has to see the same script across packs, but some overlap is often inevitable and improves the linking. Judges are asked to rank the scripts in each pack from best to worst, in terms of a holistic judgment of quality of performance, making allowance for any perceived differences in difficulty of the two tests. Tied rankings are discouraged. Judges have access to the question papers and mark schemes in order to inform their judgments.

The ranked data are converted to paired comparison data prior to analysis with the Rasch model. In our research we have used FACETS (Linacre, 2006a) to estimate the parameters. The analysis estimates a scale value (the 'measure') for every script on a single scale of perceived quality. Scripts that won or lost all their comparisons (from being ranked top or bottom in every pack in which they appeared) receive a measure based on extrapolation. These extrapolated measures are checked for plausibility. The quality of the final constructed scale is evaluated in the usual way, by considering indices of separation reliability and fit. This shows the extent to which the judges had a shared conception of the trait – in other words, the extent to which they agreed about which scripts were better and which were worse.

The final outcome of the exercise is a graph which plots the script mark against the script measure for both tests separately, as in Figure 4. The regression lines summarise the relationship between mark and measure and thus allow equivalent marks (in the sense of corresponding to the same perceived measure) on the two tests to be identified. For example, in Figure 4 if the cut-score on test A was at a mark of 25, then the corresponding cut-score on test B would be approximately 28.

Importantly, the correlation between mark and measure on each test indicates the extent to which the trait perceived by the judges was related to the trait underlying the raw scores on the tests. It is in theory entirely possible for these correlations to take any value from -1 to +1 because the mark totals play no part in either the judgments or the model fitting. Obviously, the stronger the correlation, the greater the validity of the exercise. The data could fit the Rasch model very well and produce a reliable scale of perceived quality, and yet the script measures could correlate poorly with the mark totals.

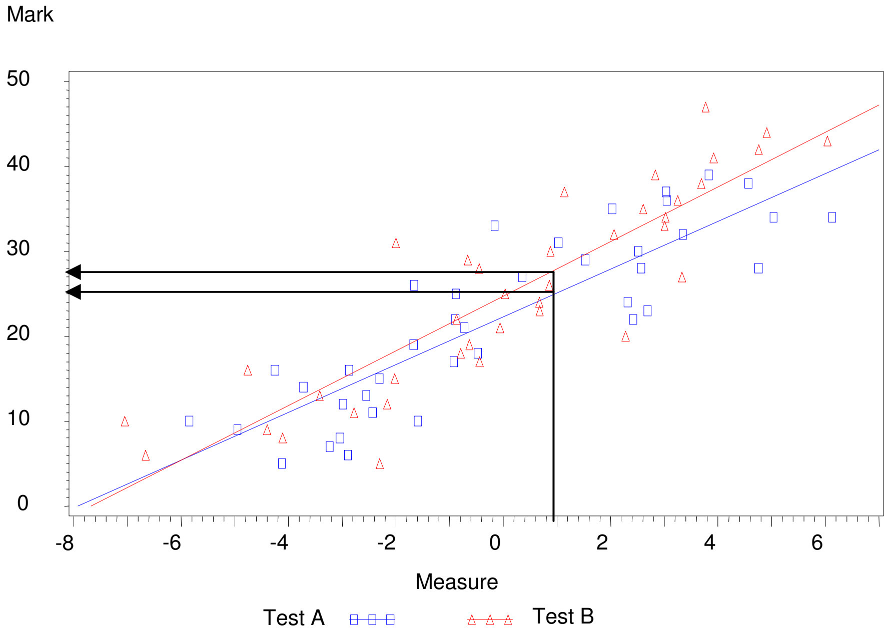

Figure 4: Illustration of standard maintaining using the rank-ordering method.

Appraisal of the rank-ordering method

Advantages / benefits

Two desirable features of the rank-ordering method arise directly from its roots in paired comparisons:

1. A natural task for the judges
The relative judgments involved in paired comparisons or rankings are easy to explain to the judges, and easy for them to carry out, in contrast to the kind of judgments involved in some standard setting exercises. For example, the well-known Angoff standard-setting method requires the judges to form a concept of a 'minimally competent candidate', and then to make a judgment about the probability of success for such a candidate on each question on the test. These kinds of judgment are relatively unfamiliar in everyday life, whereas making a choice between two alternatives, or putting things into an order of preference or merit is a much more familiar activity.

2. The judges' internal standards cancel out
This is the main advantage of using relative judgments. If the judges had to make absolute judgments about whether the scripts were above or below the performance standard then differences among the judges in where they thought the standard was located on the latent trait (see Figure 1) would affect the result. With relative judgments, the judges' individual standards are eliminated experimentally (see Andrich, 1978; Bramley, 2008 in press).

Three further desirable features of the method arise from analysing the data in the framework of an explicit measurement model (in our research we have used the Rasch model):

3. The model handles missing data
If the data fit the Case 5 Thurstone / Rasch model, the estimate of the separation between any two scripts does not depend on which other scripts they are compared with. This means that data can be missing in a non-random way without affecting the results. However, the precision (standard error) of each script's estimate depends largely on how many comparisons it has been involved in, so some effort needs to be made to ensure that each script is compared a similar number of times in total.

4. Fitting an explicit model gives insight into the data
Using the Rasch approach to analysing Thurstone pairs data means that the outcome of each individual comparison is explicitly modelled. It is therefore possible to generate a 'residual' at the individual comparison level. This residual is the difference between the observed outcome (1 or 0 corresponding to win or loss) and the expected outcome (calculated from equation (4) using the estimated script parameters). The diagnostic potential of analysing these residuals has been highlighted by Pollitt (1999; 2004; Pollitt & Elliott, 2003). They can be standardised and aggregated in various ways and used to investigate different questions of interest. For example, residuals can be aggregated for each judge to investigate judge misfit. Judges with a high value of misfit have tended to make more 'surprising' judgments than the other judges, that is, they have occasionally (or frequently) rated a low script above a high one on the scale. This can be interpreted as indicating that this judge has a different conception from the other judges about what makes a script better or worse.
Similarly, residuals can be aggregated for each script to indicate the extent to which the judges agreed on its location in the scale. Scripts with a high value of misfit have been involved in more surprising judgments than the other scripts and thus might be investigated to see if they contain anything unusual in terms of candidate performance.
Finally, at a more mundane level, individual comparisons with a large residual might indicate an error by the judge in recording their judgment on the record sheet, or an error in data entry prior to analysis.

5. Opportunities for validation
The rank-ordering method is a 'strong' method because it can highlight invalidity in two distinct ways. First, it is possible that the ranking/paired comparison data does not fit the model, and/or does not create a meaningful scale. This would indicate that the judges did not perceive the trait in the same way, or that their rankings were little different from chance. Second, as mentioned above, it is possible that the scale of perceived quality does not agree with the raw mark scale. Either of these calls the validity of the final linking into question. (A cynic might see this as a disadvantage of the method!)

Two further advantages arise from using rank ordering as opposed to paired comparisons:

6. Time saving
When the objects being compared are examination scripts, which take time to read, the paired comparison process can be extremely tedious and time-consuming, making it an arduous task for the judges. When judges have been asked for feedback after taking part in an exercise, this aspect is one that usually gets mentioned (Bramley, 2008 in press). Putting ten scripts into rank order takes much less time than would the 45 paired comparisons which can be derived from a ranking of ten objects. Thurstone (1931) used this approach of extracting every possible paired comparison from rank-ordered data for exactly this reason – to make the task less laborious.

7. Covering the whole mark range
A consequence of the time saving is that the judges can make judgments about scripts covering the whole effective mark range of the tests, rather than having to focus on scripts at or around the performance standard. This allows *all* scripts to contribute to the best-fit lines summarising the relationship between mark and measure (Figure 4). It is clear from inspection of Figure 4 that over a small mark range there is a very low correlation between mark and measure. Our research carried out to date has often shown that within a judge's pack the correlation between rank order and mark is low, in fact sometimes negative – yet when the results are aggregated over the entire mark range for all judges the overall correlation between mark and measure is high (around 0.8 to 0.9).

Disadvantages / criticisms

1. Psychological validity
This criticism applies to both paired comparisons and rank-ordering in this context. The objects being judged (scripts) are complex and the time taken to read each script is significant, making the judgmental task a slow and serial one, in contrast to the kind of judgments made in Thurstone's work, which could be made quickly and in parallel. (Examples of judgments in Thurstone's work included judges deciding which of two statements they agreed with more, or which of two handwriting samples they thought was better, or which of two statements reflected a more positive attitude towards religion). Arguably the slow, serial nature of the ranking judgments means that the 'discriminal processes' evoked by the first scripts in a pack have to be re-evoked when later scripts are compared with them. Therefore there is an element of recall involved, and it is possible that features of the later script might 'interfere' with the memory of the earlier scripts. This raises the possibility that the order in which scripts are read might affect the ranking. Thus far we have not made any attempt to control or investigate order effects, although given the variety of possible orderings within any ranking this would not be easy. We have physically arranged the scripts in a random order within each pack when assembling them, so if judges read the scripts in the order they appear we might hope that any order effects would cancel out on average.

A second criticism concerns the ability of the judges to allow for differences in difficulty of the tests being compared. How can better performance on easier questions be compared with worse performance on harder questions? In traditional test equating the performance on common items allows this linking. Use of the rank-ordering method implies that the judges can somehow perceive trait location directly. Although this criticism also applies to other methods that use expert judgment in the absence of common item linking, it is still necessary to understand the factors (both internal to the judge and external to the scripts) that cause them to make their judgments.

2. Violations of local independence
It is immediately clear that creating a set of paired comparisons out of a ranking violates the assumption of local independence between pairs. For example, if script A is ranked first, script B second and script C third then this creates the paired comparison outcomes 'A beats B', 'B beats C' and 'A beats C'. If these three scripts were to be compared in a 'true' paired comparison exercise it would be possible to obtain the inconsistent triad 'A beats B', 'B beats C' and 'C beats A'. The ranking therefore constrains the possible set of paired comparison outcomes. The more objects that are ranked, the greater this constraint (Linacre, 2006b). However, in practice it is possible that the violations of local independence do not greatly affect the results. Bramley (2005b) showed that effectively the same set of script measures were produced by analysing the rankings with the Rasch Partial Credit Model (Wright & Masters, 1982) as with the paired comparison model, but the latter appeared to create a more discriminating scale – that is, the separation (reliability) indices in the paired comparison analysis were artificially inflated. There is scope for further experimental investigation of the difference between measures created from rankings analysed as paired comparisons and measures created from a genuine paired comparison design. For example, it may be that if the objects to be ranked are sufficiently far apart on the psychological scale then many of the theoretically possible inconsistent triads would be so unlikely as to have effectively a zero probability, making the constraint imposed by a ranking much less in practice than it seems in theory.

3. Use of the Rasch model
The Rasch model used to analyse the data is analogous to Thurstone's case 5 model, which makes the questionable assumption that the discriminal dispersions of all the objects are equal. Thurstone clearly did not expect this to hold for any but the most simple stimuli. The scripts used in rank-ordering exercises are obviously far more complex than any used by Thurstone, so it seems naïve to expect this assumption to hold here.

Inspection of equations (2) and (4) above shows that allowing discriminal dispersions in Thurstone's model to vary would be equivalent to allowing the scaling parameter a in the logistic form of the model to vary. This parameter is known as the 'discrimination' parameter in IRT modelling. It is inversely proportional to σ_{AB} in Thurstone's model – that is, the smaller the discriminal dispersions, the greater the discrimination, which makes intuitive sense.

The question of whether it is justifiable to use Thurstone's Case 5 law is thus analogous to the (much debated) question of whether it is justifiable to use a 1-parameter (Rasch) model rather than an IRT model containing more parameters. If the link with Thurstone's psychophysical theory is removed altogether and the analysis of the data is treated as a statistical modelling problem then multilevel logistic models have been suggested as the most appropriate choice (Goldstein, 2008 in press).

We have preferred to use the Rasch model partly because the philosophy of Thurstone and Rasch (constructing psychological scales capable of yielding sample-free measurement) seems to be directly relevant to the issue of maintaining performance standards, but also for the pragmatic reason that the Rasch model is likely to be more robust and appropriate with the relatively small data sets produced by a rank ordering exercise. Furthermore, once the reliability of the constructed scale has been verified with separation indices, and misfitting data removed (if necessary), it seems unlikely that using a more complex model would substantively alter the estimated measures of the scripts. However, this is an issue that can be explored in future research.

4. Design and preparation

At a practical level, the design of a rank-ordering study is not a trivial task. Care must be taken to ensure sufficient linking throughout the data matrix in order to be able to estimate all the parameters within an unambiguous frame of reference (Linacre, 2005). It is also desirable to aim for each script to appear in a similar number of packs, for judges not to encounter the same script too often across packs, and for each pack to contain an appropriate range of scripts in terms of trait location in order to maximise information. We are currently experimenting with some algorithms for allocating scripts to packs that provide a good starting point for a design.

The method is also labour intensive in terms of cleaning mark totals from scripts and copying them. The former is becoming less of a problem in the UK as examination boards move towards on-screen marking of scanned images of scripts – 'clean' copies of scripts can be printed from the scanned images. Ideally the whole rank-ordering exercise would be handled digitally without using paper. The work of Richard Kimbell and colleagues (Kimbell, 2007; Kimbell et al., 2007) is showing what is possible here. A truly flexible on-line system could allow the design to be modified as the study is happening, with on-the-fly calibration guiding the selection of scripts into packs to be presented to each judge (Pollitt, 2004).

Applications of the rank-ordering method

Maintaining standards in parallel tests

Here 'parallel' means tests of the same subject, constructed to the same test specification and intended to be of approximately similar difficulty, but with no common items that would allow standard statistical equating methods to be applied.

The rank-ordering method presented here was first applied in the context of standard maintaining in national tests of English for 14-year olds in England, where the aim was find the cut-scores on the mark scale of the test under development that corresponded to the cut-scores from the previous year's live test. This work is described in Bramley (2005b).

More recent work has taken place in the context of UK school examinations at GCSE and A level, but in an experimental research setting outside the live processing of the examinations. Again the aim has been to map cut-scores ('grade boundaries') from one examination session to another, but here our focus has been on trying to show that the rank-ordering method is a more valid way of using the skills of expert judges than the existing method. This work is described in Black & Bramley (in press) and Gill et al. (submitted).

The method for capturing expert judgment currently mandated by the regulatory procedures (QCA, 2007) is known as 'top-down bottom-up'. The expert panel of judges familiarise themselves with the performance standard at the cut-score by studying 'archive scripts' from previous examination sessions that were exactly on the grade boundary. They then scrutinise scripts covering a range of marks (typically 5 to 10) in the part of the raw score scale where the grade boundary is likely to lie. This range is chosen based on the Principal Examiner's (PE's) recommendation, and statistical information about the cumulative percentage of the cohort at each mark point. The judgments are made in a particular order – starting from the top of the range of marks, awarders determine the lowest mark for which there is consensus that the work is worthy of the higher grade. This mark is the top of the 'zone'. Then, working from the bottom up, they determine the highest mark for which there is consensus the work is not worthy of the higher grade. The mark above this is the bottom of the 'zone'. The results are summarised in a tick chart (figure 5)

	Judge						
Mark	1	2	3	4	5	6	7
31	✓	✓	✓	✓	✓	✓	✓
30	✓	✓	✓	✓	✓	✓	✓
29	✗	?	✓	✓	?	✗	✓
28	?	✓	✓	✗	?	✓	?
27	✓	✓	✗	✓	✗	✗	✗
26	✗	✗	?	✗	✗	?	✗
25	✗	✗	✗	✗	✗	✗	✗

Figure 5: Illustration of tick chart and 'zone' from a mark of 26 to 30.

The expert judges then use their collective professional judgment to recommend a single mark in the zone where the boundary should lie, drawing on a variety of sources of evidence such as mark distributions, entry patterns, prior attainment, performance in previous sessions, forecast grades, and monitoring and research reports.

Although this procedure has the advantages of transparency and speed compared to rank-ordering, we have argued (Black & Bramley, in press) that rank-ordering is a more valid way to use the experts' judgment, for the following reasons:

- The ranking judgments are ‘purer’ than the top-down bottom-up judgments because they take no account of external statistical information, or of the mark totals on the scripts. They are based purely on a holistic judgment of relative quality of performance;
- The internal standards of the judges are eliminated as described previously;
- The judges work independently, so there is no opportunity for peer pressure or other social factors to influence judgments.
- The judges see more scripts in total, and scripts covering the whole mark range (rather than just around the grade boundaries). This allows the relationship between the raw mark scales and the latent trait of perceived quality to be quantified, as in Figure 4. It is this quantification that lies at the heart of the rank-ordering method.

Can the rank-ordering method be applied to any kind of test? Given that the judgments are holistic judgments of quality, it would seem to be more applicable to tests where the marking (scoring) of items also requires a more holistic judgment – that is, to tests containing essay questions or other complete performances (works of art, music, design etc). We have used the method successfully on components of assessments requiring a few longer written responses (GCSE English) and on components requiring a larger number of shorter written responses (A level Psychology), but have not yet tried to use it on tests requiring very constrained responses, or numerical answers.

Setting equivalent cut-scores on tests of different difficulty (‘vertical equating’)
Some assessment situations require equivalent cut-scores to be set on tests of the same construct that are intentionally designed to be of different difficulty. For example, some GCSE examinations have ‘tiered’ question papers. In English GCSE, the higher and foundation papers have the following overlap of grades available:

Table 1: overlapping grades on higher and foundation tiers.

Tier								
Higher	A*	A	B	C	D	E		
Foundation				C	D	E	F	G

In this case, as well as ensuring that standards are maintained within a tier from session to session, it is also necessary to ensure that the cut-scores for the overlapping grades reflect the same location on the latent trait. The rank-ordering method is an appropriate judgmental method for trying to achieve this – but the judges have to make a much greater allowance for differences in test difficulty when comparing performances from the higher and lower tier. We have investigated this once (in a research exercise separate from the ‘live’ grading of the examination) and found that the judges were able to carry out the task and create a meaningful scale. The mapping of the raw scores corresponding to the critical grade ‘C’ boundary was close to that achieved by the live processing of the exam. (Black & Bramley, in preparation).

Investigating comparability between non-parallel assessments
Here ‘non-parallel’ means tests of the same subject area, but produced to different test specifications or from different syllabuses or curricula. This situation arises in England when comparing assessments of the same subject at the same level from different examination boards. So far, the paired comparison method has been the favoured approach, but Bramley (2008 in press) has suggested that rank-ordering could offer some significant improvements.

This scenario also arises internationally – for example in ‘aligning’ state standards in the US or Australia, or even in making comparisons between countries. To do so moves further from Thurstone’s original theory, but as long as there is some justification for believing that judges can form a shared conception of a common trait in non-parallel assessments, and can perceive relative differences on this trait, then the rank-ordering method is a potential means of investigating comparability or alignment.

Validation of the rank-ordering method

It is tempting to ask whether rank-ordering gives the 'right answer' – but in the situations where we have recommended using the method, there is no well-defined right answer. (If there were, then the method that produced it would obviously be the one to use!) We would therefore argue that the rank-ordering method should not be evaluated purely on whether it can replicate the results of an existing method. For example, we have found that the rank-ordering method, when applied in a research setting, has sometimes given similar or identical results to those arrived at in a GCSE or A level award meeting – but that it has also sometimes given different results. We would recommend that the results produced by the method be combined with other relevant evidence when making a final decision about the location of a cut-score – but that this combination should ideally be of *independent* pieces of evidence, and that the values and expectations of the decision-makers be articulated when deciding what weight to give the different pieces of evidence. A possible way of doing this using Bayes' theorem was sketched out by Bramley (2006).

Our validation research on the rank-ordering method is therefore now focussing on the stability and replicability of the outcome of a rank-ordering study (as illustrated by the type of graph shown in Figure 4) when factors incidental to the theory behind the method are altered:

The setting of the exercise
Black & Bramley (in press) found that the outcome of a study was replicated when the exercise was carried out postally (i.e. with judges working alone at home) compared to when all the judges were together in a face-to-face meeting.

The features of the design
Bramley & Gill (in preparation) used existing data sets from previous rank-ordering studies and investigated the effect of systematically removing judges and scripts, and reducing the amount of linking across packs. We found that the analysis was quite robust with respect to removing judges – the separation index and the correlation between measure and mark remained high, but as more judges were removed the final result (the linking from one mark scale to the other) became more unstable. Similarly, removing scripts from packs did not affect the properties of the scale too drastically, but did have more impact on the final linking. Reducing the overlap between the packs gave an unacceptable result due to low mark-measure correlation. This idea of overlap between packs is analogous to common-person or common-item equating, where the strength of the linking is increased by increasing the number of common elements.

Determining the optimum number of scripts to include in a pack involves trading off the time savings from having more scripts (yielding more paired comparisons) against the cognitive complexity of the task for the judges. In a recent study Black (in preparation) found that judges reported that compiling a ranking of three scripts was easier than ranking ten, even though the three were much closer together in terms of location on the latent trait than the ten were.

Summarising the relationship between mark and measure
The choice of a linear regression of mark on measure to summarise the relationship between the mark scale and the measure scale is somewhat arbitrary. Bramley & Gill (in preparation) investigated the effect of using a variety of other summary best-fit lines: measure on mark regression, Standardised Major Axis (SMA), information-weighted regression, non-linear (loess), and non-linear (local linear smoothing). Reassuringly, the different methods all produced similar outcomes, although there were some notable differences for cut-scores nearer the extremes of the mark scale (which are more influenced by fluctuations in the shape of the line).

We discovered that the issues involved in choosing a best-fit line to summarise the relationship between two variables are more complex than might at first appear (see Warton et al. 2006 for a review), and concluded that there was no compelling reason to stop using the mark-on-measure regression that has been used in studies carried out to date.

Quantifying the uncertainty in the outcome

It is usually possible to derive an estimate of 'equating error' from standard statistical equating methods. The equivalent of equating error in the rank-ordering method would be some indicator of the variability of the cut-score on test Y corresponding to a given cut-score on test X. In the more complex statistical equating scenarios, bootstrap resampling methods (e.g. Efron & Tibshirani, 1993) are a good choice for obtaining an estimate of the equating error (Petersen et al. 1989). Bramley & Gill (in preparation) showed that a bootstrap method can be successfully applied to the rank-ordering analysis to obtain an analogous 'margin of error' (illustrated in Figure 6 below). However, it is important to be careful in the interpretation of these margins of error. They do not answer questions about what might have happened with a different experimental design. They only relate to sampling variability in the regression lines relating mark to measure. In other words, the bootstrapping procedure treats the pairs of values (mark, measure) for each script as random samples from a population and shows what other regression lines might have been possible with other random samples from the same population.

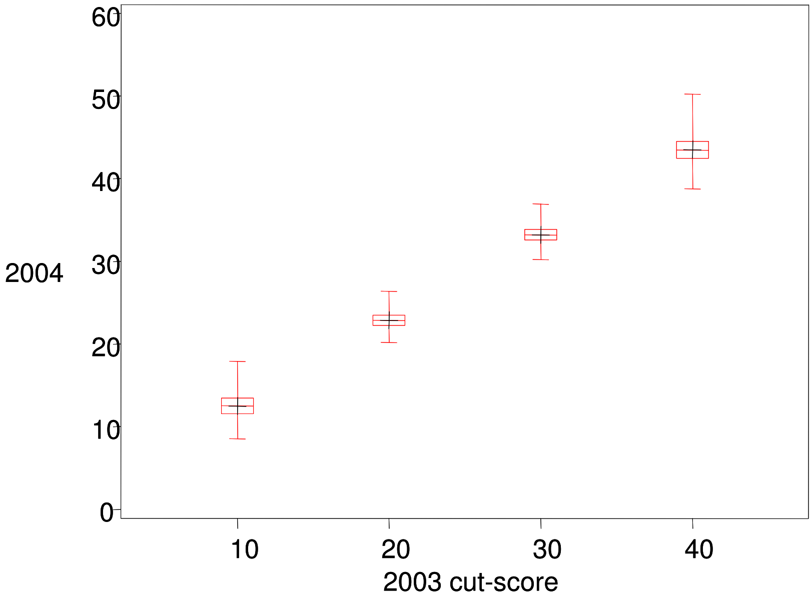

Figure 6: Bootstrap sampling variability of cut-scores on the 2004 test corresponding to cut-scores of 10, 20, 30 and 40 marks on the 2003 test (using data from Bramley, 2005b).

Bramley & Gill (op. cit.) found that there was more sampling variability in the intercept and slope of the line when there was less of a linear relationship between mark and measure. This fits with intuition – we would expect to be less confident in the outcome of an exercise where the scale of perceived quality created by the judges bore less relation to the actual marks on the scripts.

The equating error can of course be reduced by increasing the sample size (i.e. the number of scripts in the study). Bramley & Gill found that retrospectively halving the sample size in the study shown in Figure 6 increased the equating error obtained from bootstrapping by up to 60%. The actual amount varied across the four cut-scores. There was more variability for cut-scores at the extremes of the mark range than for those in the middle, as might be expected, since in general, predictions from regression lines become less secure as the lines are extrapolated further from the bulk of the data. This suggests that ranges of scripts should be chosen for the study which ensure that the key cut-points (e.g. grade boundaries) to be mapped from one test to another do not occur at the extremes.

Conclusion

The rank-ordering method is grounded in the psychometric theory of Thurstone and Rasch. It is a good choice of method for linking raw scores on one scale to raw scores on another where:

- Standard statistical equating methods are not possible because there are no common items or people;
- The products of the assessment (i.e. the examinees' work) are suitable for global, holistic judgments of quality
- It is believed that expert judges have a shared conception of what is better and worse and are capable of making valid relative judgments of quality of performance, allowing for differences in task difficulty.

The method has been successfully applied to linking score scales on parallel tests, and to linking score scales on tests designed to be of different difficulty. There is great potential for the method to be applied to the investigation of comparability or alignment issues of assessments in the same subject area produced by different agencies.

References

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

Andrich, D. (1978) Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement* 2, 449-460.

Black, B. (in preparation). *Investigating January versus June awarding standards using an adapted rank-ordering method.*

Black, B., & Bramley, T. (in press). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education*.

Black, B. & Bramley, T. (in preparation). *Using expert judgment to link mark scales on different tiers of a GCSE English examination: a rank ordering method*.

Bradley, R.A., & Terry, M.E. (1952). The rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrica, 39,* 324-345.

Bramley, T. (2005a). Accessibility, easiness and standards. *Educational Research, 47(2)*, 251-261.

Bramley, T. (2005b). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement,* 6 (2) 202-223.

Bramley, T. (2006). *Equating methods used in Key Stage 3 Science and English.* Paper for the NAA technical seminar, Oxford, March 2006.

Bramley, T. (2008 in press). Paired comparison methods. In P.E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Bramley, T. & Gill, T. (in preparation). The rank-ordering method for standard maintaining: an investigation of factors affecting the stability of the outcome.

Cizek, G.J. (2001). *Setting Performance Standards: Concepts, Methods and Perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. Boca Raton, Florida: Chapman & Hall/CRC.

Gill, T., Bramley, T., & Black, B. (submitted). An investigation of standard maintaining in GCSE English using a rank-ordering method.

Goldstein, H. (2008 in press). Commentary on statistical issues arising from chapters. In P.E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards.* London: Qualifications and Curriculum Authority.

Hambleton, R.K., & Pitoniak, M.J. (2007). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement.* 4th Edition. (pp. 433-470). ACE/Praeger series on higher education.

Holland, P.W., & Dorans, N.J. (2007). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement.* 4th Edition. (pp. 187-220). ACE/Praeger series on higher education.

Kimbell, R. (2007, October). *Technology and the assessment of creative performance.* Keynote presentation at the Cambridge Assessment conference, Cambridge. Available at http://www.assessnet.org.uk/file.php?file=/1/Resources/Conference_2007/Richard_Kimbell_Paper.pdf (Accessed 14/01/08).

Kimbell, R., Wheeler, T., Miller, S., & Pollitt, A. (2007) *E-scape portfolio assessment phase 2 report*. (2007). London: Goldsmiths, University of London. Available at http://www.goldsmiths.ac.uk/teru/UserFiles/File/e-scape2.pdf (Accessed 14/01/08)

Kolen, M.J., & Brennan, R.L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices.* (2nd ed.). New York: Springer.

Linacre, J. M. (2005). A User's Guide to FACETS Rasch-model computer programs. www.winsteps.com.

Linacre, J.M. (2006a). FACETS [Computer program, version 3.60.0]. www.winsteps.com

Linacre, J.M. (2006b). Rasch analysis of rank-ordered data. *Journal of Applied Measurement, 7(1)*, 129-139.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming and equating. In R. L. Linn (Ed.), *Educational Measurement.* 3rd Edition. (pp. 221-261). Phoenix, Arizona: The Oryx Press.

Pollitt, A. (1999, November). *Thurstone and Rasch – Assumptions in scale construction*. Paper presented at a seminar held by the Research Committee of the Joint Council for General Qualifications, Manchester. Reported in B.E. Jones (Ed.), (2000), *A review of the methodologies of recent comparability studies*. Report on a seminar for boards' staff hosted by the Assessment and Qualifications Alliance, Manchester.

Pollitt, A., & Elliott, G. (2003). *Monitoring and investigating comparability: A proper role for human judgement*. Cambridge: Research and Evaluation Division, University of Cambridge Local Examinations Syndicate.

Pollitt, A. (2004, June). *Let's stop marking exams*. Paper presented at the annual conference of the International Association for Educational Assessment, Philadelphia.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

QCA (2007). *GCSE, GCE, VCE, GNVQ and AEA Code of Practice*. London: QCA.

Thurstone, L.L. (1927a). Psychophysical analysis. *American Journal of Psychology, 38*, 368-389.

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

Thurstone, L.L. (1931). Rank order as a psychophysical method. *Journal of Experimental Psychology, 14,* 187-201.

Warton, D.I., Wright, I.J., Falster, D.S., & Westoby, M. (2006). Bivariate line fitting methods for allometry. *Biological Reviews, 81,* 259-291.

Wright, B.D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14(2)*, 97-116.

Wright, B.D., & Bell, S.R. (1984). Item banks: what, why, how? *Journal of Educational Measurement, 21(4)*, 331-345.

Wright, B.D., & Masters, G.N. (1982). *Rating Scale Analysis.* Chicago: MESA Press.