



CAMBRIDGE ASSESSMENT

Evaluating the rank-ordering method for standard maintaining

Tom Bramley, Tim Gill & Beth Black

Paper presented at the International Association for Educational Assessment
annual conference, Cambridge, September 2008.

Research Division
Cambridge Assessment
1, Regent Street
Cambridge
CB2 1GG

Bramley.T@cambridgeassessment.org.uk

Gill.T@cambridgeassessment.org.uk

Black.B@cambridgeassessment.org.uk

www.cambridgeassessment.org.uk

Cambridge Assessment is the brand name of the University of Cambridge Local Examinations Syndicate, a department of the University of Cambridge. Cambridge Assessment is a not-for-profit organisation.



UNIVERSITY of CAMBRIDGE
Local Examinations Syndicate

Evaluating the rank-ordering method for standard maintaining

Abstract

The rank-ordering method for standard-maintaining was designed for the purpose of mapping a known cut-score (e.g. a grade boundary mark) on one test to an equivalent point on the mark scale of another test, using holistic expert judgments about the quality of exemplars of candidates' work (scripts). How should a method like this be evaluated? If the correct mapping were known, then the outcome of a rank-ordering exercise could be compared against that. However, in the contexts for which the method was designed, there is no 'right answer'. This paper presents an evaluation of the rank-ordering method in terms of its rationale, its psychological validity, and the stability of the outcome when various factors incidental to the method are varied (e.g. the number of judges, the number of scripts to be ranked, methods of data modelling and analysis).

1. Introduction

A considerable amount of research at Cambridge Assessment has investigated developing the rank-ordering method as a method of standard maintaining – mapping scores on one test to equivalent scores on another test via expert judgment. The details of the method are described in Bramley (2005) and Bramley & Black (2008). Its specific application to standard maintaining in the context of public examinations in England (GCSE, AS and A-level¹) is described in Black & Bramley (2008). The theory underlying the method (Thurstone's law of comparative judgment) is described in detail in Bramley (2007).

The purpose of this paper is to show how a new method such as the rank-ordering method can be evaluated. The main focus of validation is to show that the method does what it claims to do. The remainder of this introduction gives a brief overview of the rank-ordering method for standard maintaining. The subsequent sections describe the validation research we have carried out.

In a rank-ordering exercise judges are given packs containing scripts, cleaned of marks, from two or more tests, and asked to place the scripts in each pack into a single rank order of perceived quality from best to worst, making allowances for differences in difficulty of the question papers. The rank orders are converted into sets of paired comparisons and then analysed with a Rasch formulation of Thurstone's (1927) paired comparison model, as described in Bramley (2007). The outcome of this analysis is a 'measure' of perceived quality, on a single scale, for each script in the study. The distance between two scripts on the scale is related to the probability of one being judged to be better than another in a paired comparison. A plot of the script total marks against these measures allows the mark scales of the tests to be compared, as shown in Figure 1 below.

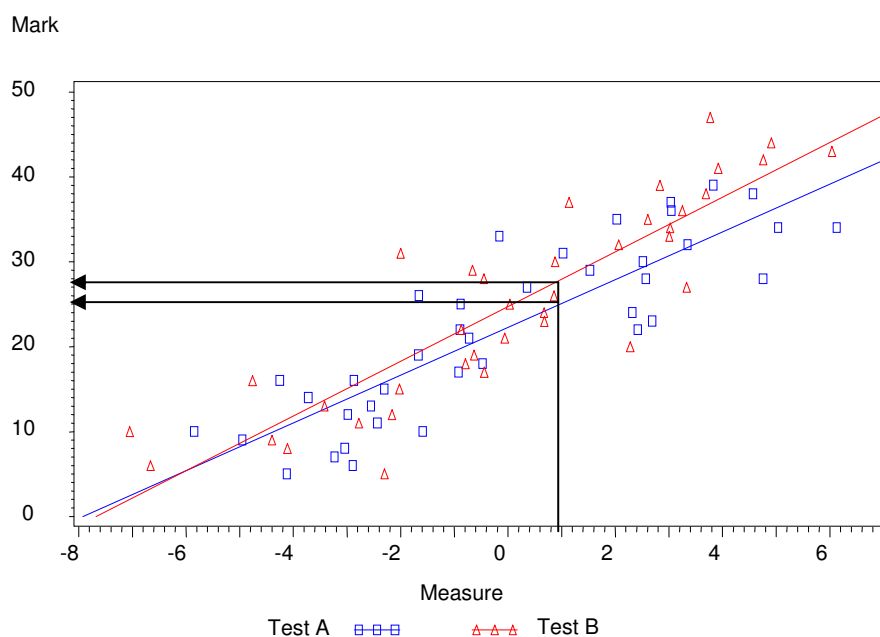


Figure 1: Example of a plot of mark against measure for two tests.

¹ GCSE = General Certificate of Secondary Education. It is the standard subject qualification taken by the majority of 16 year-olds in England and Wales. A level GCE = Advanced Level General Certificate of Education. These qualifications typically require two years of study beyond GCSE, with the first year of work being assessed at Advanced Subsidiary (AS) level.

For example, Figure 1 shows that a mark of 25 on Test A corresponds to the same measure as a mark of 28 on Test B. Thus if 25 were a grade boundary mark on Test A then the equivalent grade boundary on Test B, according to this method, would be 28.

2. Validating the method

The rank-ordering method claims to map a score on one test to the equivalent score on another test. It is thus making the same kind of claim as methods of statistical test equating. These divide into two main groups – those based on classical test theory and those based on latent trait theory (see Kolen & Brennan, 2004, for details). The rank-ordering method has the same conceptual foundation as the latent trait statistical equating methods. The tests to be equated are assumed to measure the same trait – that is, it is assumed that all the candidates (whichever test they have taken) can be located along the same abstract continuum. The direction of causality is taken to be from trait to test scores – that is, differences in trait level among the candidates cause the differences in the test scores (Borsboom, 2005). The same raw scores on two different tests do not necessarily imply the same trait level (henceforth ‘ability’) because the test items could differ in difficulty. The purpose of equating is to find the raw score on one test that implies the same ability as any given raw score on the other test.

The usual way to estimate the abilities of all candidates on the same scale is to have a common element link between the two tests to be equated – occasionally a common person link, more often a common item link. In the situations for which the rank-ordering method was designed, neither of these linking methods is possible. This can be for a number of reasons:

- The tests are not considered suitable for application of standard latent trait 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).

2.1 Psychological validity

The assumption underlying the rank-ordering method is that expert judges are able to provide the common element link for latent trait equating. This implies that they can estimate the ability of a candidate, given the test questions and the candidate’s answers to those questions. How they manage to do this is something of a mystery – as Bramley (2007) has pointed out, it is unlikely that they are mentally performing the same kind of iterative maximum likelihood procedure as latent trait software. Thus understanding the factors influencing their judgments is an important strand of validating the rank-ordering method. This work is in progress.

However, even though the actual cognitive processing of the judges is not well understood, the rank-ordering method was designed to maximise the validity of the judgmental process, by requiring the judges to make relative, rather than absolute judgments. The advantages of this have been well rehearsed elsewhere (e.g. Pollitt, 1999; Bramley, 2007; Bramley & Black, 2008), the most important one being that differences among the judges in absolute terms (i.e. where on the trait they perceive each script to lie) cancel out, and all judges contribute to estimating the relative locations of the scripts.

2.2 Testability

A desirable feature of any method is that it should be open to validation. That is, it should be possible to assess the extent to which it has worked. This is a key feature of the rank-ordering method. The foundational assumption of the method is that differences on the same latent trait cause both differences in the test scores and differences in the expert judgments. This implies that the test scores should be correlated with the measures estimated from the expert judgments, as in Figure 1. The size of the correlations between test scores and measures is thus an indicator of the extent to which the method has been successful in any particular application.

A second way in which the results of using the method can be validated is in terms of the properties of the scale of 'perceived quality' created by the judges. This involves carrying out the usual investigations of separation reliability and model fit that are involved in any latent trait analysis (e.g. Wright & Masters, 1982; Bond & Fox, 2007).

The rank-ordering method is thus a 'strong' method, in that it can 'go wrong' in two different ways. The judges can fail to create a meaningful scale (e.g. with low separation reliability we would conclude that their judged differences between the scripts were mostly due to chance); but even if they create a meaningful scale the estimated measures can fail to correlate with the test scores – indicating that they were perceiving a different trait to the one underlying the test scores.

So, there are indicators that can tell us if the method has gone wrong – that is, if there are reasons to doubt, or give less weight to, the results. But can anything tell us that the method has produced the 'right answer'? This depends on whether one believes that there is a right answer. If an analogy with temperature measurement were justifiable, the question is like asking whether a new instrument for measuring temperature (e.g. based on electrical resistance) gives the same reading as an established mercury thermometer. If there was an established method that gave the right answer in the area of standard maintaining, we would use it! Arguably, the right answer in the situation for which the rank-ordering method was designed is in principle the result that would be obtained from a conventional statistical method carried out in ideal conditions – for example with large randomly equivalent groups taking each test. But the reason that different methods (including those involving expert judgment) are required is that those ideal conditions can not be obtained in practice in some contexts (e.g. public examinations in England), for the reasons mentioned above.

Therefore, the focus of validation has to be on the stability and replicability of the outcome of a rank-ordering study (as illustrated by the type of graph shown in Figure 1) when factors incidental to the theory behind the method are altered. If the outcome were shown to depend to an undesirable extent on artefacts of the method, then this would count against its validity. For example, Black & Bramley (2008) 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 following sub-sections of this paper report some investigations carried out on existing data sets. Section 2.3 investigates whether the strong positive correlation between mark and measure shown in Figure 1 is to some extent an artefact of the rank-ordering method and thus whether the method appears to exaggerate the extent to which the expert judges can perceive differences between the scripts.

Section 2.4 describes some analyses of existing data sets, varying some of what might be called the 'design parameters' of a rank ordering study – for example reducing the number of judges, the number of scripts in a pack, and the overlap of scripts across packs. The purpose of these analyses was to investigate how stable the final outcome of the study (the linking of the two mark scales) was with respect to this variation.

One concern about this methodology that is sometimes voiced by those encountering it for the first time is that the pack design itself may impose a natural order on the scripts and it is this that creates at least some of the correlation between mark and measure. For example, Baird & Dhillon (2005) claim that:

“Even if the examiners had randomly ordered the scripts within these packs, fairly high correlations would have been obtained once the two sets were combined”. (Page 27).

If this is true, then the overall high correlation between mark and measure illustrated in Figure 1 is an artefact, and the method exaggerates the extent to which the expert judges can perceive differences between the scripts.

The simplest way to test this claim is to take an existing data set, replace the judges’ actual rankings with random rankings, re-estimate the script measures, and plot the marks against the new estimated measures. This was carried out using the data from Black & Bramley (2008), repeating the random rankings five times with different random seeds.

The correlations between mark and measure produced by this process are shown in Table 2.1, as are the correlations from the original rank-ordering. There are two correlations for each of the five simulated rankings, one for the ‘Test A’ scripts and one for the ‘Test B’ scripts:

Table 2.1: Correlation between mark and measure in packs with ‘random’ rankings.

	Test A	Test B
Original rank order exercise	0.89	0.92
Random 1	-0.20	0.01
Random 2	-0.03	-0.29*
Random 3	0.06	-0.24
Random 4	0.14	-0.09
Random 5	0.24	-0.09

* correlation statistically significantly different from zero.

It is clear from Table 2.1 that the levels of correlation from the random rankings were very low, particularly compared with the correlations from the original research. Only one of the correlations was significantly different from zero, and this was negative, which is the opposite of what would be expected if the pack design was artificially inflating the correlation.

A plot of mark against measure for one of the simulated rankings is shown in Figure 2.2². It is clear from this plot that there is no spurious correlation between mark and measure arising from the rank-ordering method.

² We would like to thank our colleague Nat Johnson for carrying out this analysis.

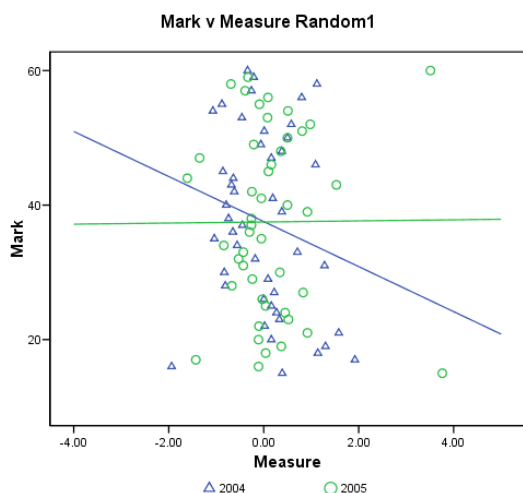


Figure 2.2: Plot of mark against measure using random rankings (including regression lines).

The reason why the correlation is not an artefact can be understood by analogy with tailored testing or adaptive testing. The design illustrated in Figure 2.1 is the same as a tailored testing design where the rows would be examinees and the columns would be test items. A more able examinee would receive a more difficult set of items than a less able examinee, and by targeting the test carefully across the ability range each examinee would only need to take a small subset of the total item pool in order to obtain an accurate ability estimate. Because of the linking from the overlapping subsets of items, each examinee obtains an ability estimate on the same scale, even though the weaker candidates never attempt the more difficult items, and the stronger candidates never attempt the easier items.

The connection of this to the rank-ordering method can be seen from the equations used to model the data. The judges' rankings are converted to paired comparisons, and the Rasch formulation of Thurstone's Case V model is used (Andrich, 1978). This expresses the probability of script A 'beating' script B (being ranked above script B) as:

$$\text{Log odds (A>B)} = \beta_A - \beta_B \approx \ln [f(\text{A>B}) / f(\text{B>A})] \quad (1)$$

where β is the location of the script on the latent trait, and $f(\text{A>B})$ is the frequency with which A is ranked higher than B.

In words: when the data fit the model, the log of the odds of success depends only on the distance between the two scripts on the latent trait. This distance is estimated by the log of the ratio of wins to losses. Similarly, the probability of script B beating script C is:

$$\text{Log odds (B>C)} = \beta_B - \beta_C \approx \ln [f(\text{B>C}) / f(\text{C>B})]. \quad (2)$$

If we add equation (2) to equation (1) we get:

$$\text{Log odds (A>B)} + \text{Log odds (B>C)} = \beta_A - \beta_B + (\beta_B - \beta_C) = \beta_A - \beta_C = \text{Log odds (A>C)}.$$

In other words, the scale separation between A and C can be estimated via the comparison with B without them having to be directly compared. As long as no script wins (or loses) all its comparisons (which in the rank-ordering method means that a script must never be ranked top or bottom of every pack in which it appears) it is possible to estimate the location of each script on the same latent trait. Having established such a scale (and confirmed its validity through the usual statistical investigations of parameter convergence, separation reliability and fit), the size of the correlation between mark and measure is a purely empirical question. It is free to take

any value in the range of the correlation coefficient (i.e. between -1 and 1). There is no *a priori* reason to expect any artificial inflation of the correlation arising from the experimental design.

2.4 Effect of ‘design features’ on the outcome of a rank-ordering study

The organiser of a rank-ordering study has to take a number of decisions about the design of their study. These include:

- the number of scripts to use
- the criteria for selecting and/or excluding scripts from the study
- the range of marks on the raw mark scale to cover
- the number of judges to involve
- the number of judgments to require each judge to make
- the number of packs of scripts to allocate to each judge
- the number of scripts from each test to include in each pack
- the range of marks to be covered by the scripts in each pack
- the amount of overlap in the ranges of marks covered by the scripts across packs
- whether to impose any extra constraints, e.g. to minimise exposure of the same script to the same judge across packs; or to ensure that each script appears the same number of times across packs.

Some of these decisions will be affected by external constraints such as the amount of money available to pay participants, the number of suitably qualified judges, the time available for the judgments, and the number of scripts available. Others are more at the discretion of the researcher.

Obviously it is not possible using existing data to investigate the effect on the outcome of increasing numbers of scripts / packs / judges because that would require extra data collection. However, it is possible to investigate the effect of reducing these numbers by selectively ‘degrading’ an existing data set. The purpose of the analyses reported in this section was to see how robust the outcomes of the analysis would be to different ways of degrading the data. The outcomes investigated were the scale properties such as separation³ and reliability⁴, the correlation between mark and measure, and the effect on the substantive result (the mark difference at the grade boundaries).

The re-analysis varied the following three factors:

1. Number of judges – it is possible in FACETS⁵ to specify which comparisons should be included in the analysis. Thus it is possible to omit all comparisons by a particular judge or judges and observe the impact on the results.
2. Number of scripts per pack – similarly it is possible to exclude all comparisons involving specific scripts in specific packs.
3. Overlap between packs – by removing all comparisons involving certain scripts it is possible to reduce the amount of overlap between packs.

The data used for these analyses came from the GCSE English rank-ordering study, reported in Gill *et al.* (in prep.). This study used a relatively high number of judges (7) and packs (6). This gave greater scope for degrading the data by reducing either of these facets.

Tables 2.2 to 2.4 below summarise the impact of removing data in the three ways described above on several aspects of the results: the separation and reliability indices, the correlation between mark and measure, and the linked grade boundaries generated. The ‘subset connection’ refers to whether measures for the scripts could be estimated on a single scale. Problems here indicate a failure of the design to link all the scripts together.

³ The separation index indicates the spread of the estimated measures for the scripts in relation to their precision (standard error).

⁴ Separation reliability is an estimate of the ratio of true variance to observed variance, calculated from the separation index.

⁵ FACETS software for many-facet Rasch analysis, Linacre (2005).

Table 2.2: Summary of results of removing judges from the analysis.

Judges removed	Subset connection	Comparisons	Separation	Reliability	Correlation		Boundaries		
					2004	2005	D	C	A
None	OK	3712	7.94	0.98	0.90	0.94	22	33	57
1 under-fitting	OK	3172	7.92	0.98	0.91	0.92	21	33	56
1 best-fitting	OK	3240	7.25	0.98	0.89	0.94	21	33	58
1 over-fitting	OK	3172	7.49	0.98	0.92	0.93	21	32	56
2 under-fitting	OK	2632	8.04	0.98	0.89	0.89	22	33	56
2 best-fitting	OK	2700	6.88	0.98	0.88	0.93	22	33	58
2 over-fitting	OK	2632	6.33	0.98	0.90	0.93	21	32	56
3 judges	OK	2160	6.14	0.97	0.89	0.91	18	31	57
3 other judges	OK	2092	6.20	0.97	0.90	0.93	23	34	56
4 judges	OK	1552	6.04	0.97	0.91	0.92	20	31	55
4 other judges	OK	1620	5.39	0.97	0.89	0.92	19	32	58
5 judges	No	1080							

Table 2.2 shows that removing more than half the judges (four out of seven) still produced an outcome with a reasonable amount of separation and reliability between the scripts' measures, as well as very high correlations between mark and measure. However, the outcome in terms of the grade boundaries⁶ was different, depending on which judges were excluded from the analysis. With up to two judges excluded, none of the three boundaries was more than one mark different from the boundary obtained in the original analysis. When three or four judges were excluded, some very different outcomes were obtained, particularly at grade D, where there were fewer scripts in the study. Removing five judges meant there were problems with the connectivity between the scripts: FACETS reported that there may have been two 'disjoint subsets', and so it was not able to estimate measures for all the scripts in the same frame of reference.

Table 2.3: Summary of results of removing scripts at random from the analysis.

Scripts removed	Subset connection	Comparisons	Separation	Reliability	Correlation		Boundaries		
					2004	2005	D	C	A
None	OK	3712	7.94	0.98	0.90	0.94	22	33	57
1 per year per pack	OK	2352	5.74	0.97	0.89	0.93	26	35	54
2 per year per pack	OK	1260	3.90	0.94	0.89	0.90	22	32	55
3 per year per pack	OK	504	1.46	0.68					

Table 2.3 shows that removing up to 2 scripts per year per pack had little adverse impact on the internal consistency of the scale, but did generate different grade boundaries. Removing 3 scripts per year per pack reduced the number of scripts per pack to 4, and hence only 6 paired comparisons per pack. Remarkably, FACETS was still able to estimate measures for most of the scripts, despite the fact that over 85% of the original data had been lost! 13 scripts were excluded out of 80, seven having won all their comparisons and six having lost theirs. The separation fell substantially to 1.46 and the reliability to 0.68. This indicates that a substantial amount of the variability among the estimated script measures was due to random error, and realistically this outcome would not be acceptable.

⁶ In GCSE, AS and A level examinations, the cut-scores are known as 'grade boundaries'. For GCSE, outcomes are reported on a grade range of A* to G. The grade boundaries at A, C, D, and F are set judgmentally while the other grade boundaries are calculated arithmetically by interpolation.

Table 2.4: Summary of results of removing overlapping scripts from the analysis.

Overlap scripts removed	Subset connection	Comparisons	Separation	Reliability	Correlation		Boundaries		
					2004	2005	D	C	A
None	OK	3712	7.94	0.98	0.90	0.94	22	33	57
One	OK	2576	6.31	0.98	0.89	0.93	20	32	57
Two	OK	1624	4.73	0.96	0.53	0.60			

Table 2.4 shows that it is important to retain a substantial amount of overlap between the packs. Removing two overlap scripts led to much lower correlations of measure with mark, 0.53 in 2004 and 0.60 in 2005, and therefore the overall results would not be acceptable. The implication is that in fact the linking was weakened too much for the analysis to cope, even if FACETS was able to find enough data to link the top packs to the rest of the packs. The measures for the top-marked scripts came out close to the overall average (zero) - suggesting that these were effectively acting as a 'disjoint subset' even though they had not been diagnosed as such by the software.

The implication of this re-analysis is that future rank-ordering studies could possibly use fewer judges or fewer scripts per pack (around 25% fewer data points than were used in the original study analysed here) and still give a valid result. However, there is no guarantee of this, since other factors not manipulated here could also be relevant.

It is also worth stressing that while this result might have a financial implication (i.e. the possibility of collecting less data with a corresponding reduction in cost), we do not know what (if any) psychological implications there might be from reducing the number of scripts per pack. The data used in this exercise all came from a real study which used ten scripts per pack. It is possible that a real study using fewer scripts per pack might raise issues not apparent from *post-hoc* manipulation of the data from a study with more scripts per pack. A recent rank-ordering study by Black (2008) used only three scripts per pack, i.e. as close as ranking can get to genuine paired comparisons. Judges reported finding the task of ranking three scripts cognitively easier than ranking ten scripts, even though those scripts were all at the same nominal standard, and came from three different examination sessions (as opposed to the two sessions in the data used here).

2.5 How should the relationship between mark and measure be summarised?

The rank-ordering technique produces a measure of 'perceived quality' for each script, and then relates this to the original mark using a regression of mark on measure, as illustrated previously in Figure 1. By having a regression line for more than one test, equivalent marks for each test can be determined graphically via the regression lines, or calculated via the regression equations, and the standard maintained in terms of the grade boundaries.

However, the choice of least squares regression of mark on measure is somewhat arbitrary, because it is not clear which of mark and measure is best described as a dependent variable or an independent variable. To what extent do the results of a rank-ordering study depend on this arbitrary choice of best-fit line?

In turning to the research literature for guidance on how to choose a best-fit line it soon became clear that we had unwittingly opened something of a can of worms. This is an issue which has been discussed in diverse fields from fishery (e.g. Ricker, 1973) to astronomy (e.g. Isobe *et al.*, 1990)! The interested reader is referred to Warton *et al.* (2006) for a review of the debates, and a clarification of the different terminology used by different researchers.

This section of the report investigates the impact of changing the type of best-fit line on the substantive result (the mark difference at the grade boundaries). This was done for five different

sets of rank order data: three from studies using Psychology A-level, and two from a study using English GCSE.

Five different types of best fit line were tried (see the appendix for an illustration of each type):

1. Regression of measure on mark - this is generating a line of best fit based on minimising the sum of the horizontal, rather than the vertical, distances between each point and the line in a plot such as Figure 1. There is no reason not to use this method as we are merely using the regression equation to provide an equivalent mark for each measure, or an equivalent measure for each mark: we are not suggesting that one somehow causes the other, nor is our focus primarily on statistical prediction of one from the other.
2. Standardised Major Axis (SMA) - sometimes referred to as a 'structural regression line', or a 'reduced major axis', or a 'type II regression' (see Warton *et al.* 2006). It has the advantage of symmetry: the best-fit line is the same whichever variable is chosen as X or Y.
3. Information weighted regression - under the Rasch model, each estimated measure for the scripts has an associated standard error, which indicates how precisely the measure has been estimated. Thus it is possible to give more weight to the more precise measures in the (Y|X) regression. To do this we used a WLS (Weighted Least Squares) regression, where the weight was the 'information' - the inverse of the square of the standard error. Thus a measure with a lower standard error was given a higher weight in determining the slope of the best-fit line.
4. Loess best fit line - a non-linear regression line that takes each data point (or a specified percentage of points) and fits a polynomial to a subset of the data around that point. Hence it takes more account of local variations in the data.
5. Local Linear Regression Smoother. This is very similar to the Loess regression line in that it divides up the line into small chunks and only uses data from that part of it to calculate the regression.

Table 2.5 Impact of alternative methods for summarising the mark v measure relationship.

Psychology AS unit	Data set 1		Data set 2		Data set 3	
Grade boundary ⁷	E	A	E	A	E	A
Regression of mark on measure	25	48	25	47	29	45
Regression of measure on mark	24	48	25	47	30	45
Standardised Major Axis	25	48	25	47	29	45
Information Weighted Regression	25	48	25	48	28	45
Loess Smoother	26	49	25	47	28	45
LLR Smoother	26	48	26	48	28	45

English GCSE unit	Data set 1		Data set 2		
Grade boundary	F	C	D	C	A
Regression of mark on measure	17	38	22	33	57
Regression of measure on mark	15	39	24	35	56
Standardised Major Axis	16	39	23	34	57
Information Weighted Regression	17	40	20	32	56
Loess Smoother	19	43	24	32	56
LLR Smoother	19	43	25	31	58

Table 2.5 shows that the effect of changing the method of summarising the relationship between mark and measure was generally not large – a reassuring finding. This was particularly the case for the Psychology studies, where the differences were not more than two marks. There were larger differences when looking at the English studies and these were mostly at the lower grade boundaries (F and D). It may be that this was due to there being less data around these marks,

⁷ For AS and A levels, outcomes are reported on a grade range of A to E (pass) and U (fail). The grade boundaries for A and E are set judgmentally. The intervening grade boundaries are calculated arithmetically by interpolation.

meaning that one or two scripts could have had a larger influence on the regression line, pulling it in one direction or the other. The Psychology studies had more data around the lower grade boundary and were thus more stable.

Are there any grounds for deciding which is the best method to use? As there is no particular reason to regress mark on measure as opposed to measure on mark, it may be that the SMA line is more appropriate than either of those two. Warton *et al.* (2006) state that subjects (i.e. scripts in our case) should be randomly sampled when fitting a SMA line, which is not what happens in a rank-ordering study where scripts are sampled to cover the mark range fairly uniformly. However, since the scripts for both years are sampled in the same way, and the interest is in the comparison from one test to another rather than the actual value of the parameters of the regression line, this might not be a serious problem. Probably a more practical drawback is the relative difficulty of calculating the SMA line, compared to a Y|X regression line which is readily available in all standard software. A final criticism which could be levelled at the SMA line (e.g. Isobe *et al.* 1990) is that its slope does not depend at all on the correlation between the two variables. Again, this is arguably not a major problem, since assessing the relation between the two variables can be done either by inspecting the graph or via the usual correlation coefficient. As discussed previously, this correlation is an important aspect of the validity of the exercise – the more the judges perceive the trait of ‘quality’ differently from how the mark scheme awarded marks, then the lower will be the correlation between mark and measure and the less confident we can be that the exercise was relevant to standard maintaining.

It may be that using the Loess or LLR regression lines better summarises the relationship between mark and measure, if this relationship is not linear. On the other hand, these lines are more difficult to understand than the simpler least squares regression line, and cannot be summarised in one simple equation. They are also less robust to variability in the data – i.e. require more data points to be estimated reliably. Another drawback of using the Loess and LLR regression lines applies to the context of standard maintaining in GCSEs and A-levels - because these lines are not constrained to be straight, they are less likely to produce intermediate grade boundaries which are equally far apart. Current awarding procedures (QCA⁸, 2007) require the difference between each judgmentally determined pair of grade boundaries to be split into equal intervals when determining the arithmetic boundaries (allowing one mark difference if the division is not exact).

The information weighted regression has some intuitive appeal, but arguably if we take account of error in the measures, we should also take account of error in the marks.

Taking all these considerations together, it seems as though there is not as yet any very convincing reason to shift from the Y|X regression of mark on measure which has been used in rank-ordering studies to date.

2.6 How can uncertainty in the outcome be quantified?

A desirable feature of any method is that it should be possible to quantify the amount of uncertainty in the outcome. In the rank-ordering context this question is difficult to address, because of the number of variables – what if we had used different judges? Or different scripts? Or a different experimental design for allocating scripts to packs and packs to judges?

However, it is clear that the outcome does depend on the regression equations summarising the relationship between mark and measure illustrated in Figure 1. Section 2.5 has shown that there was little effect of using different ways of summarising the relationship between mark and measure in the Psychology studies, but slightly more effect with the GCSE English. Intuitively we would expect that the weaker the relationship between judged measure and mark (as

⁸ Qualifications and Curriculum Authority – the regulator of qualifications in England.

expressed, for example, by the correlation or by the standard error of estimate), the less confidence we might have in the outcome of the exercise. This is because a weaker relationship suggests that the judges were perceiving qualities in the scripts which were different from those qualities which led to the marks awarded.

However, the standard error of estimate (the RMSE) of the regression is not the appropriate measure of confidence in the outcome. It is simply the root mean square of the residuals (vertical distance from point to regression line). It therefore indicates how far on average each point is from the line. The purpose of the rank-ordering exercises is to estimate the two regression lines in order to compare *average* differences between the two tests. It is therefore the variability in the intercepts and slopes of these lines that is of interest in quantifying the variability of the outcome.

This effect of this variability can most easily be quantified using a bootstrap resampling process (Efron & Tibshirani, 1993). Effectively a large number of random samples of size N (where N is the number of scripts from a particular year) are drawn (with replacement), and the slope and intercept parameters of the regression of mark on measure are estimated for each sample. Then, for any particular mark on one test, the distribution of the equivalent mark on the other test can be obtained.

Data from two rank-ordering exercises, one using data from Key Stage 3 English⁹ reported in Bramley (2005) and one using data from AS Psychology reported in Black & Bramley (2008) were used for the bootstrap resampling.¹⁰ Table 2.6 below shows the equation of the regression lines of mark on measure from these studies.

Table 2.6: Regressions of mark on measure in two rank-ordering studies.

	Test	Slope (s.e.)	Intercept (s.e.)	RMSE	R ²
KS3 English	2003	2.96 (0.18)	23.54 (0.57)	3.38	0.88
	2004	3.06 (0.18)	26.49 (0.58)	3.29	0.90
AS Psychology	2003	4.03 (0.37)	38.00 (1.16)	6.94	0.77
	2004	4.08 (0.40)	37.81 (1.10)	7.47	0.70

1000 re-samples were taken from each year's test within each study and values for the intercept and slope parameters obtained. Then, using the procedure described above, for each set of values for the slope and intercept parameters, the mark on one version of the test corresponding to a particular mark on the other version was calculated. In the KS3 English, where the test total was 50 marks, the distribution of marks on the 2004 test corresponding to marks of 10, 20, 30 and 40 on the 2003 test was obtained. In the AS Psychology, where the test total was 60 marks, the distribution of marks on the 2004 test corresponding to marks of 20, 30, 40 and 50 on the 2003 test was obtained.

The results are shown in figures 2.3 and 2.4 on the next page.

⁹ Key Stage 3 tests are taken by pupils at age 14 in England.

¹⁰ The SAS macro %boot was used for the resampling in this report. This macro is available for download from the SAS Institute website: <http://support.sas.com/ctx/samples/index.jsp?sid=479&tab=details> Accessed 15/02/07.

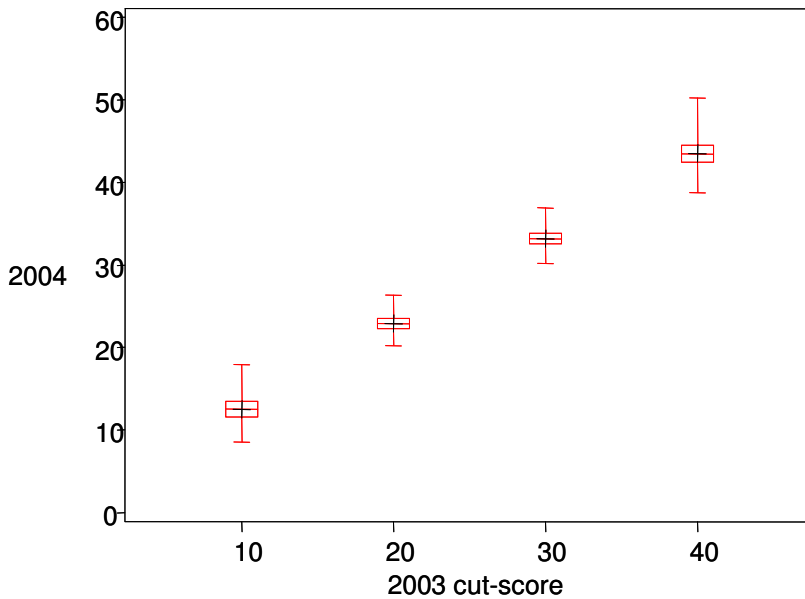


Figure 2.3: KS3 English – distribution of scores in 2004 corresponding to scores of 10, 20, 30 and 40 in 2003.

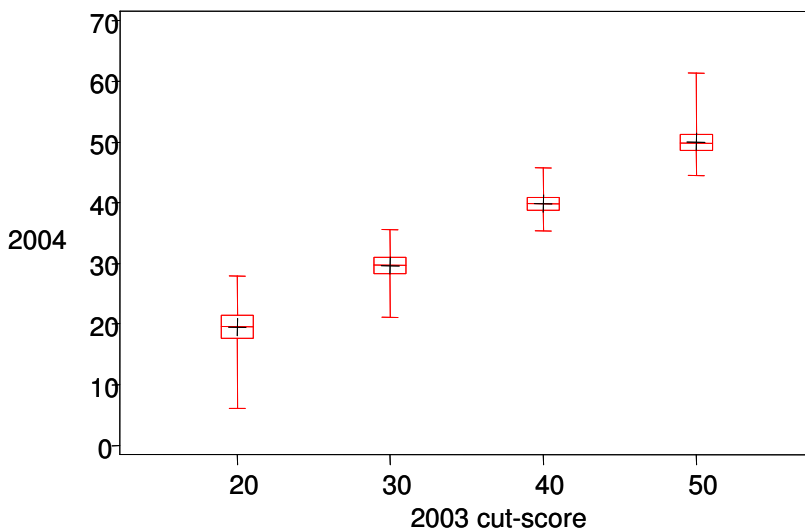


Figure 2.4: AS Psychology – distribution of scores in 2004 corresponding to scores of 20, 30, 40 and 50 in 2003.

Several observations can be made from the information in the table and boxplots:

- The interquartile ranges (IQRs) show that the middle 50% of scores corresponding to a given mark fell in a relatively narrow range – about 2 marks for the KS3 English, and about 3 marks for the AS Psychology.
- There was more variability at all four cut-scores in the AS Psychology, as would be expected from the worse fit of the regression line.
- In both studies, the variability was less for the two cut-scores in the middle of the mark range than at either extreme.
- The full range of the distributions (min to max) could be quite wide, particularly at the extremes (over 21 marks in the case of AS Psychology at the 20-mark cut-score).

Bootstrap resampling thus seems to be a good way of quantifying the uncertainty in the outcome of a rank-ordering exercise. 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.

There is more sampling variability in the intercept and slope of the line when there is less of a linear relationship between mark and measure. (If the relationship were perfectly linear then all the re-sampled datasets would produce the same regression line). This makes sense – 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 sampling variability can of course be reduced by increasing the sample size (i.e. the number of scripts in the study). It was not possible to retrospectively increase the sample size for the studies reported here, but halving the sample size in the KS3 English study increased the IQR of the distributions (the width of the 'box' in the boxplots) by up to 60%. The actual amount varied across the four cut-scores. This suggests that it might be possible to estimate in advance of the study the number of scripts needed to achieve a desired level of stability in the outcome.

The fact that there was more variability for cut-scores at the extremes of the mark range than for those in the middle was also to be expected. 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.

Finally, the possibility of quantifying the uncertainty in the rank-ordering outcome allows information from other sources in a standard-maintaining exercise (e.g. score distributions, knowledge of prior attainment) to be combined in a principled way. An outline of how this might be done using a Bayesian approach was sketched in Bramley (2006). This is an interesting area for future investigation.

3. Summary and conclusions

A considerable amount of research has been devoted to exploring the scope of, and validating, the rank-ordering method. The aim has been to make explicit the assumptions underlying the method, and to discover the extent to which the outcome (the mapping of scores on one test to another) is affected by varying various incidental features of the method.

In summary:

- The rank-ordering method has the same theoretical rationale as latent trait test equating methods.
- The main assumption is that expert judges can directly perceive differences in trait location among the objects (scripts) being judged, allowing for differences in difficulty of the test questions.
- The relative (rather than absolute) judgments made in a rank-ordering exercise do not require all the judges to share the same absolute concept of a given performance standard.
- The validity of the judgments made, in terms of the features of scripts influencing the judgments, is currently being investigated.
- The method is a 'strong' method, in that it is possible to evaluate the extent to which it has worked in a given situation. There are two independent ways in which it can fail – the judges can fail to create a meaningful scale, and the scale they create can fail to correlate with the mark scale.
- The correlation between mark and measure that emerges from the latent trait analysis is not an artefact of the design – random rankings lead to correlations close to zero.

- The method is robust in that outcomes are fairly constant when factors such as the setting of the exercise, the number of judges, the number of scripts per pack, and the method of plotting the best-fit line are varied.
- The uncertainty in the outcome can be quantified by bootstrap resampling.

Whether the rank-ordering method becomes the preferred method in any particular standard-maintaining context is likely to depend on a number of factors. Chief among these is likely to be the level of satisfaction with the existing method. For example, Black & Bramley (2008) have argued that the rank-ordering method would be a more valid source of evidence for decision-making in a GCSE or A-level award meeting¹¹ than the existing ‘top-down bottom-up’ method (see QCA (2007) for details of this process). Of course, in practice other factors will need to be traded off against validity in deciding on the most appropriate method – for example cost, transparency, scalability and flexibility. However, we feel that just because a method is well established and familiar (such as the ‘top-down bottom-up’ method) this should not exempt those promoting it from explicating its rationale and providing evidence of its validity, along the same lines as those presented here.

¹¹ The award meeting is the mandated process by which grade boundaries are decided using a combination of expert judgement and statistics.

References

- Andrich, D. (1978) Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement* 2, 449-460.
- Baird, J.-A., & Dhillon, D. (2005). *Qualitative expert judgements on examination standards: valid, but inexact*. Guildford: AQA. RPA_05_JB_RP_077.
- Black, B. (2008). *Using an adapted rank-ordering method to investigate January versus June awarding standards*. Paper presented at EARLI/Northumbria Assessment Conference, Berlin.
- Black, B., & Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education* 23 (3).
- Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Borsboom, D. (2005). *Measuring the mind: conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.
- Bramley, T. (2005). 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. Available at http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Conference_Papers Accessed 3/07/08.
- Bramley, T. (2007). Paired comparison methods. In P.E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. (pp. 246-294). London: Qualifications and Curriculum Authority.
- Bramley, T., & Black, B. (2008). *Maintaining performance standards: aligning raw score scales on different tests via a latent trait created by rank-ordering examinees' work*. Paper presented at the Third International Rasch Measurement conference, University of Western Australia, Perth, January 2008.
- Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.
- Gill, T., Bramley, T., & Black, B. (in prep.). An investigation of standard maintaining in GCSE English using a rank-ordering method.
- Isobe, T., Feigelson, E.D., Akritas, M.G., & Babu, G.J. (1990). Linear regression in astronomy. I. *The Astrophysical Journal*, 364, 104-113.
- 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.
- Pollitt, A. (1999). *Thurstone and Rasch – assumptions in scale construction*. Awarding Bodies' seminar on comparability methods, held at AQA, Manchester.
- Qualifications and Curriculum Authority (QCA). (2007). GCSE, GCE, GNVQ and AEA Code of Practice. London, QCA. Available from

<https://orderline.qca.org.uk/bookstore.asp?FO=1169415&Action=PDFDownload&ProductID=9999078664> Accessed 3/07/08.

Ricker, W.E. (1973). Linear regressions in fishery research. *Journal of the Fisheries Research Board of Canada*, 30, 409-434.

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

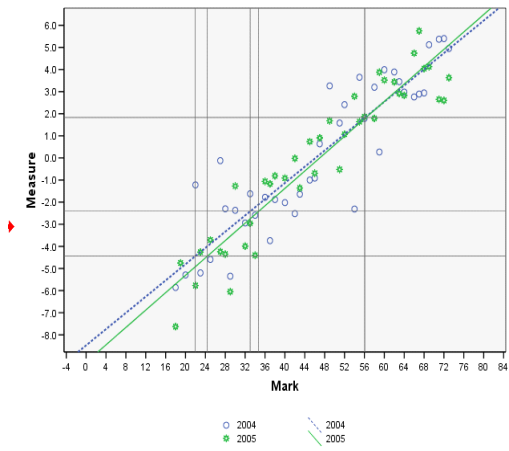
Warton, D.I., Wright, I.J., Falster, D.S., & Westoby, M. (2006). Bivariate line fitting methods for allometry. *Biological Reviews*, 81, 259-291.

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

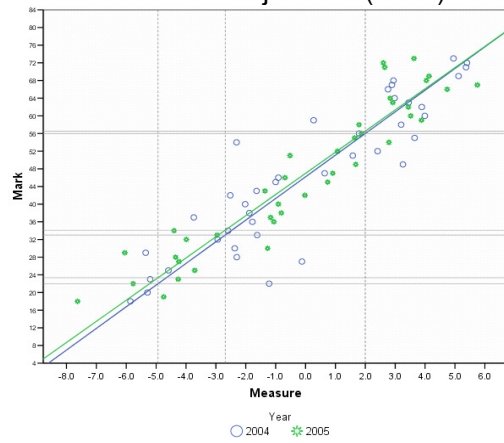
Appendix

Examples of plots with different best fit lines.

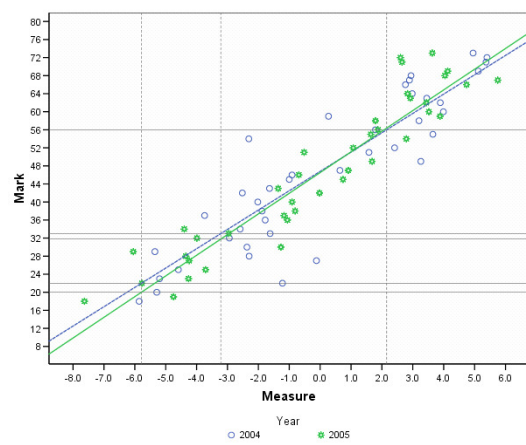
1. Measure on mark regression



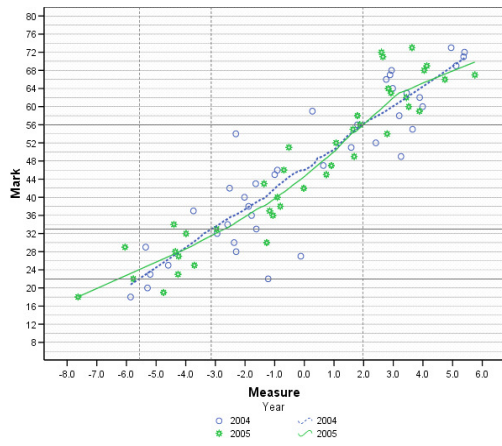
2. Standardised major axis (SMA)



3. Information weighted regression



4. Loess smoother



5. LLR smoother

