CAMBRIDGE ASSESSMENT

An investigation of standard maintaining in GCSE English using a rank-ordering method.

Tim Gill, Tom Bramley and Beth Black

Paper presented at the British Educational Research Association annual conference, Institute of Education, London, September 2007.

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

Gill.Tim@cambridgeassessment.org.uk
Bramley.T@cambridgeassessment.org.uk
Black.B@cambridgeassessment.org.uk
www.cambridgeassessment.org.uk

Abstract

This paper tests a new method of standard maintaining in UK examinations using expert judgement devised by Bramley (2005); that of rank-ordering. This method allows the raw mark scale on one test to be compared to the raw mark scale on another, and the equivalent marks determined in terms of perceived quality of performance. If the two tests are from successive years then it can be used to maintain standards between years.

The current standard maintaining practice at awarding meetings involves judges looking at scripts and comparing their quality with an internal standard (e.g. a concept of what an A grade script looks like, based on prior inspection of archive scripts from previous years). This concept may differ between examiners. The advantage of rank-ordering is that it is based on the Thurstone paired comparison technique, which involves direct comparisons of scripts and thus eliminates this internal standard. The Thurstone method has been used on many occasions in research comparing examination standards over time and between different awarding bodies.

Our research builds on previous investigations of rank-ordering as a method of standard maintaining (Bramley, 2005; Black & Bramley, 2006) which demonstrated that it has a potential role in the awarding process. These were based on question papers with mainly short answer questions. Thus, the main aim of the current research was to test the method on a paper with long essay type questions and to see if the results were comparable. The OCR GCSE unit English 2431Non-Fiction, Media and Information was chosen. The results showed that the method worked as well as in the previous research; correlations between the measures of script quality from the rank-ordering and the original marks were high, as were levels of agreement between judges on the relative quality of the scripts. The grade boundaries generated by the rank-ordering were generally a few marks lower than the original awarding decisions, but this is likely to have been due to the different information that fed in to the decisions in each case. Thus, we have further evidence that rank ordering may have a role in making awarding decisions, alongside other judgemental and statistical information.

This paper focuses on the analysis of results and interpretation of statistical output from the software, to assist other researchers and practitioners who are interested in trying out the technique.

Introduction

At the end of compulsory education in England, at age 16, pupils undertake examinations that go towards qualifications in the General Certificate of Secondary Education (GCSE). A pass in these exams is graded on a scale A*-G. The results of these are used to monitor national standards, to compare school performance in the form of league tables and for selecting those pupils going on to further education. It is therefore vital that these grades represent the same standard each year. The current method of maintaining standards is via an award meeting, where the cut-scores (or grade boundaries) are determined by a mixture of expert judgement and statistics as mandated in the Qualifications and Curriculum Authority (QCA) Code of Practice (QCA, 2006):

1) After studying archive scripts to familiarise themselves with the standards set in previous sessions, several examiners (all experts in their subject) scrutinise sample scripts on a range of marks determined prior to the award meeting and use their expert judgement to determine whether the work is worthy of achieving the grade or not. This process generally leads to the identification of a range of marks on which there is no consensus among the examiners either way (worthy of the grade or not). This forms the 'zone' within which the boundary will lie.
2) A variety of other information is then considered to help reach a final decision on the boundary mark. This includes score distributions, size and composition of the entry, forecast grades and on occasion prior attainment and relevant research reports. Equivalent information from previous sessions is also available to provide context and help interpret changes.

The first part of this process involves examiners making judgements that are essentially subjective (see Cresswell, 2000; Greatorex, 2003; Pollitt & Crisp, 2004). It will depend on their concept of what the standard should be, as well as differences in the exam papers between years, which will test an overlapping but different subset of the syllabus and may vary in difficulty. Even where archive scripts are used to compare standards between the current year and previous years, examiners are not making a direct comparison but, "a comparison of the two inferred standards, each of which is based upon an interpretation by the observer" (Cresswell, 2000, p70).

In practice, lack of space and time at award meetings means archive scripts are rarely used (Murphy *et al.*, 1996), meaning the first part of the process relies on the examiners having internalized the standards at each of the key grade boundaries. In other words, that they have a concept of what an A grade 'is' and are then able to compare a script with this concept. As Bramley (2005) notes, there are reservations about examiners using an abstract internalised standard when making their judgements. It has been shown that, psychologically speaking, humans are better at making comparative rather than absolute judgements (Laming, 2004). There are further drawbacks of the award meeting process, such as the outcome depending on the leniency or severity of the examiners present, lack of time leading to less careful decision making as well as the tendency for social dynamics to influence proceedings (Murphy *et al.,* 1996; Black & Bramley, 2006).

A judgmental technique which (in theory) eliminates the use of abstract internalised standards is Thurstone's paired comparison method (Thurstone, 1927), where pairs of objects are compared on the basis of a single attribute or trait (e.g. 'attractiveness', 'goodness', or, in the case of examination scripts, 'perceived quality of performance'). If these comparisons are repeated across judges and different pairs of objects a single scale can be constructed for the trait and each object located on the scale. In the mid 1990s the Thurstone method was adopted by assessment researchers for investigating comparability of standards in the same subject over time (Bell *et al.*, 1998) and in the same subject between different examination boards (see for example Elliott and Greatorex, 2002; Adams & Pinot de Moira, 2000; Jones & Meadows, 2004).

It has been the technique adopted by QCA for their Standards Over Time reports since 2005 (QCA, 2006). The main drawback with this approach for standard maintaining is the time consuming nature of making a large number of paired comparisons, and the tedium for the judge panel involved.

This drawback led Bramley (2005) to develop the rank-ordering method – an adaptation of the paired comparison method where, rather than carrying out repeated paired comparisons, judges rank sets (packs) of ten scripts. The data from such a ranking can be analysed as though it had come from 45 paired comparisons, but can be collected in much less time. Whereas the paired comparison method focuses judgments on particular grade boundaries, the rank-ordering method can cover the whole effective mark range. It allows the raw mark scales of two separate tests to be compared so that one can say that mark x on Test A is equivalent to mark y on Test B, in the sense of the same perceived quality of candidate performance on the test. If the two tests are the same but from successive years then the standard from the first year can be carried forward to the next year. As with the paired comparison method, rank-ordering aims to ensure that the examiners are making comparisons between objects, rather than with an internalized standard.

Each script appears in several packs and from these repeated rankings obtains a 'measure', which is its location on the latent trait scale of perceived quality. Analysis is then possible in terms of agreement between the measure and the mark as well as the 'fit' of any particular script or judge. Since each pack contains scripts from two successive years, one can also infer via the measure that a score of x in one year is equivalent to a score of y in the other year.

The method was found to work well with national tests in England at age 14 (Bramley, 2005). Correlation coefficients between the measure for each script generated by the paired comparisons and the script's original mark were high, 0.95 for both 2003 and 2004 scripts. Comparing the standards from both years using this method suggested that the 2004 test was approximately three marks easier at all levels than 2003. This agrees well with the actual cut scores, which concluded the 2004 test was two marks easier at all levels. Similarly, Black & Bramley (2006) showed the method worked for an AS level paper in Psychology, comparing standards from 2003 to 2004 and from 2004 to 2005. This produced correlations of mark and measure of between 0.81 and 0.92. There was some agreement and some discrepancy between the rank-ordering outcomes and the awarding meeting outcomes in terms of the judgementally determined cut scores. This research included a replication of the 2003/4 study, which showed that carrying out the exercise by post produced very similar outcomes to carrying the exercise out by face-to-face meeting.

Thus, the method seems robust. However, further evidence is needed to assess the appropriateness of different question formats for rank ordering. Since the Psychology AS level paper investigated by Black & Bramley (*op. cit.*) consisted of many short answer questions, the primary purpose of this research was to look at a question paper with long, essay type questions. For this we used an English GCSE paper, comparing the standards in 2004 and 2005. It may be that these types of questions lend themselves more easily to the holistic judgement required of rank ordering. If so, we should obtain results that better fit the model and are more consistent with the judgements made in an award meeting. We consider the results in terms of correlations between mark and measure, the overall fit of the model to the data, and how the cut scores generated by the rank-ordering method compare to the actual cut scores from the awarding meeting.

We also look in some depth at the rank-ordering methodology and interpretation of the output from the computer package that produces the script 'measures'.

Methodology

Script Selection

The OCR (Oxford, Cambridge and RSA Examinations) GCSE unit English 2431 (Non-Fiction, Media and Information) from specification 1900 was chosen for this research exercise. This unit has two separate exams, which target pupils of differing abilities. The foundation tier is aimed at lower ability pupils and thus only pass grades C to G are available. The higher tier targets higher ability pupils with grades A* to E available. The foundation tier of paper 2431 has a maximum mark of 60 and consists of two or three short answer questions (worth 1, 2 or 3 marks each) and three further essay type questions (worth 15-20 marks each). The higher tier paper is out of 90 and consists of three essay type questions all worth 30 marks each. These papers therefore contrast with the Psychology paper investigated by Black & Bramley (2006) which consisted entirely of short answer questions. This English unit was first examined in 2003 and has been stable in 2004 and 2005 in terms of candidature and continuity of senior examining personnel.

Both the foundation and higher tier papers were included in the research. For the foundation tier (grades C-G) a single script was selected for most, but not all, mark points between 10 and 45 from both years. Scripts scoring less than 10 were not included because performance at this level is very patchy, and this is 4 or 5 marks below the G grade boundary. 57 was the maximum mark scored on the paper, but as a mark of 45 was already 5 marks above the C grade boundary it was deemed more important to get the required 'overlap' of marks between the packs than to include scripts from the very top of the range. On the higher tier, a single script was selected for most of the mark points between 18 and 73 for both 2004 and 2005. Once again this easily covered the range of marks for all of the grade boundaries.

There were a few general principles that were followed when selecting each script.

- Since the scripts were to be photocopied, it was preferable to use scripts with handwriting that was relatively clear, and without too much annotation from markers.
- Scripts where all questions were attempted were chosen ahead of ones where some questions were not answered at all. It has been shown that inconsistent scripts (where a candidate scores highly on some questions, but poorly on others) are more difficult to grade than consistent scripts (Scharaschkin & Baird, 2000).
- Scripts marked by examiners who were scaled[1] were avoided.

All scripts were cleaned of examiners' marks and annotation (e.g. ticks, comments and other notation) as well as centre number.

Pack design

The design of the packs was such that there was enough linking of scripts between judges and packs to be able to locate all the scripts on the same scale. The design thus aimed to ensure that each script was compared with as many other scripts as possible, and also that each script was judged by several different judges.

The basic design was as follows – each judge received a set of packs of scripts that spanned the whole of the mark range. Pack 1 contained scripts at the top of this range, pack 2 scripts with slightly lower marks (but including marks overlapping with pack 1), and so on until the bottom of the range was reached. In general, each pack had a mark range of 10-15 marks and overlap with adjoining packs of 2-8 marks.

[1] Examiners who consistently over- or under-mark might have all their marks adjusted. For example, an examiner who marks with consistent mild severity might be scaled by +2, that is, adding two marks to all scripts in their allocation.

An example of how packs 1 and 2 for each judge might look and the interlinking between and within judges is shown in figure 1 below. The actual complete design is given in Appendix A.;

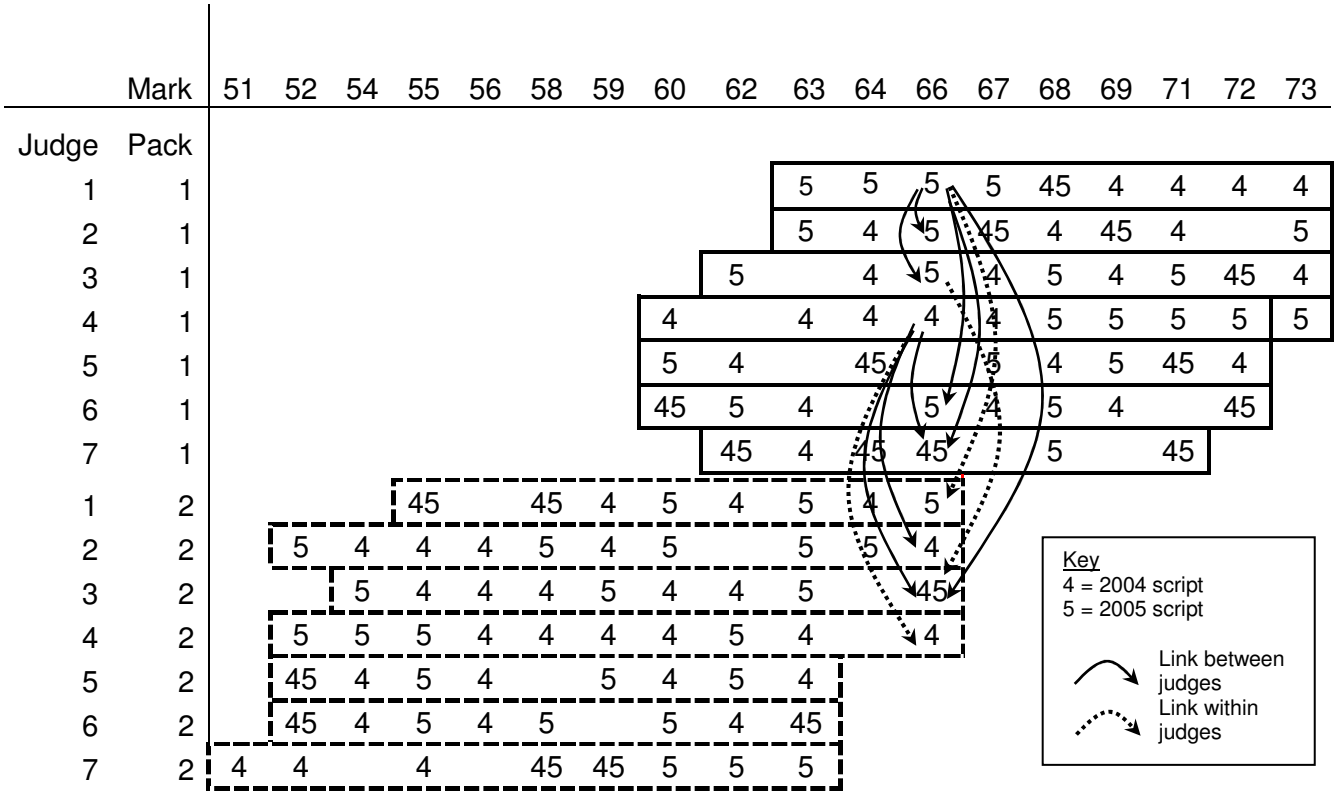


Figure 1: An example of pack design

The boxes with unbroken lines represent the mark ranges for pack 1 and the dotted lines for pack 2 for each of the seven judges. A '4' means the pack includes a script from 2004 with that mark, a '5' represents a script from 2005. The overlap in marks between packs is clear in that the scripts in pack 1 range from 60 to 73 marks and for pack 2 from 51 to 66 marks.

Figure 1 also demonstrates how the linking works (at just one mark, 66). The unbroken arrows show links between judges, the dotted arrows are links within judges. So, on mark 66, the 2005 script is seen by judges 1, 2, 3, 6 and 7 in their pack 1, and judges 1 and 3 in their pack 2. The 2004 script is seen by judges 4 and 7 in their pack 1, and by judges 2, 3 and 4 in their pack 2. This means judges 1 and 3 see the 2005 script in both packs, whilst judge 4 sees the 2004 script in both packs.

Similar linking of the scripts could be demonstrated on marks, 60, 62, 63 and 64 on figure 1, and this design was followed throughout the other packs. In conjunction with the overlapping mark ranges across packs, this meant that there were links, indirectly, between the very top script and the very bottom script.

Further features of the pack design:

- Each judge received a unique combination of scripts in each pack. However, all scripts were judged by more than one judge.
- Each set of packs contained a mixed pattern of scripts – some packs contained 2004/5 scripts from the same range of marks, others 2004 higher than 2005 and others 2005 higher than 2004.
- The design was such that the number of times a particular judge saw a particular script was minimised to avoid over-familiarity. The maximum number of times a judge saw a

particular script was four times. The majority of scripts were seen only once or twice by judges.

Allocating scripts to judges

Each judge was allocated eleven separate packs of ten scripts to be ranked. These were divided up as follows;

- Packs 1-5 contained foundation tier scripts only, with each pack containing five scripts from 2004 and five scripts from 2005.
- Packs 6-11 contained higher tier scripts only, with each pack containing five scripts from 2004 and five scripts from 2005.

The judges were sent all eleven packs in the post. Black & Bramley (2006) found that the outcomes of a rank-ordering exercise were very similar whether the exercise was carried out by post or by a face-to-face meeting, allaying any concern that the more cost-effective postal method might affect the results.

Judges

All seven members of the panel (five principal examiners and two assistant principal examiners) from the June 2005 awarding meeting agreed to take part in the study.

The task

The judges were asked to place the ten scripts within each pack into a single rank order from best to worst. They were instructed that they could use any method they wished to create their rankings, based on reading the scripts and using their own judgements to summarise their relative merits, but they should not re-mark the scripts. They were instructed to combine the scripts from two different years into a single rank order using their own experience and knowledge to make allowances for differences in the question papers. Allowing for differences in papers is a process that judges will be used to from award meetings. In order to assist this process, we supplied each judge with a question paper and mark scheme for each year. Tied rankings were discouraged.

Judges were told that the order of the scripts in each pack at outset was genuinely randomised. Additionally, they were told not to make assumptions about the relative quality of scripts from each year within a pack. Judges were provided with a record sheet for each pack, upon which they recorded their rankings.

Data analysis

For each pack, the judges had compiled a single rank order of the ten scripts from best to worst. In order to analyse the data it was necessary to convert this raw ranked data into paired comparisons. Thurstone himself sometimes did this conversion in order to save time in collecting data (Thurstone, 1931). This conversion is straightforward in that the top script on the list ‘won’ all of the paired comparisons with the nine scripts below it and the second script on the list ‘won’ the comparisons with the eight scripts below it, but was ‘beaten’ by the one script above it and so on. Every pack of ten ranked scripts thus generated 45 paired comparisons.

The paired comparison data was analysed by fitting the following Rasch model (Andrich 1978):

$$P_{ij} = \frac{e^{(B_i - B_j)}}{1 + e^{(B_i - B_j)}} \qquad (1)$$

where P_{ij} = the probability that script i beats script j in a paired comparison,
B_i = the measure for script *I,*
B_j = the measure for script *j.*

Thus, the probability that one script 'beats' another in a paired comparison is modelled as a function of the difference between the measures for each script. The measures were estimated using FACETS software (Linacre, 2005) which uses an unconditional maximum likelihood algorithm. This iteratively refines estimates of the measures until the difference between the expected number of wins and losses according to equation (1) and the observed number of wins and losses for each script falls below a pre-set value.

Fit

The standard way of assessing the level of fit of the data to the model is by examining the residuals: the difference between the expected outcome and the actual outcome for each judgement. A large residual occurs when a script with a high measure is ranked below a script with a low measure. Summing the squared residuals for each script or judge and then averaging gives the *infit mean square* and *outfit mean square* (Wright & Stone, 1979), which are measures of misfit for each judge and script[2]. An outfit or infit mean square greater than one indicates more variation between the observed and expected judgements than predicted by the model, whereas a value less than one indicates less variation than predicted. FACETS reports outfit and infit measures for each judge and script (and the standardized infit and outfit) and also lists the most misfitting judgements (i.e. the individual judgements with the largest standardized residuals). If any badly misfitting scripts or judges are identified they can be excluded from the analysis and the measures re-estimated, in order to assess the substantive impact of misfit on the outcome.

Separation and reliability

The separation index is a measure of the spread of the estimates compared to their precision (standard error). The higher this value, the greater the confidence that differences between the measures are due to genuine measured differences rather than random error. Similarly, the separation reliability is the ratio of true variance to observed variance, which indicates the proportion of the variation in the measures which can be attributed to differences between the scripts. It is analogous to Cronbach's Alpha in traditional test theory. For further details on these indices see Wright & Stone (1979).

[2] The only difference between infit and outfit is that the former is information weighted, and thus places more weight on the well-targeted observations, and less weight on the extremes. Thus, outfit is more sensitive to outliers.

Results

Model Fit

The output from the first FACETS run for the foundation tier highlighted four scripts with particularly high outfit, the source of which could be traced to one highly misfitting judgement in each case. This occurred when the script with the much higher measure was placed lower in the comparison; hence these judgements had very high residuals (standardized residuals of 9). The decision was taken to exclude these badly misfitting judgements, re-run the FACETS analysis and see if this reduced the outfit mean square to an acceptable level. The other paired comparisons involving those scripts in those packs were also excluded. Following the FACETS re-run these scripts were no longer badly misfitting. There were now no scripts or individual judgements that were as badly misfitting, suggesting a more valid outcome.

The amount of judge misfit was reasonable. After the re-run, one judge had a standardized infit of 3.0 which is higher than ideal, though by no means high enough to warrant the removal of all of that judge's data (which would have reduced the number of judgements significantly, thus reducing the reliability of the other measures).

The separation and reliability measures were very high on both runs. After excluding the misfitting scripts the separation was 8.76 and separation reliability was 0.99. We can therefore confident be that the observed differences between scripts were not due to measurement error.

In the FACETS output for the higher tier there were four scripts with very high infit or outfit, which it was deemed worth excluding from the packs where they had a particularly misfitting judgement. As with the foundation tier, in each case this was due to a script with a much higher measure being ranked below a script with a lower measure.

Once again the judge misfit was reasonable, with judge 3 being the only slight concern (standardized infit = 2.2, outfit = 2.3).The separation was 7.94 and reliability 0.98. We are therefore confident that the observed differences between scripts are not due to measurement error.

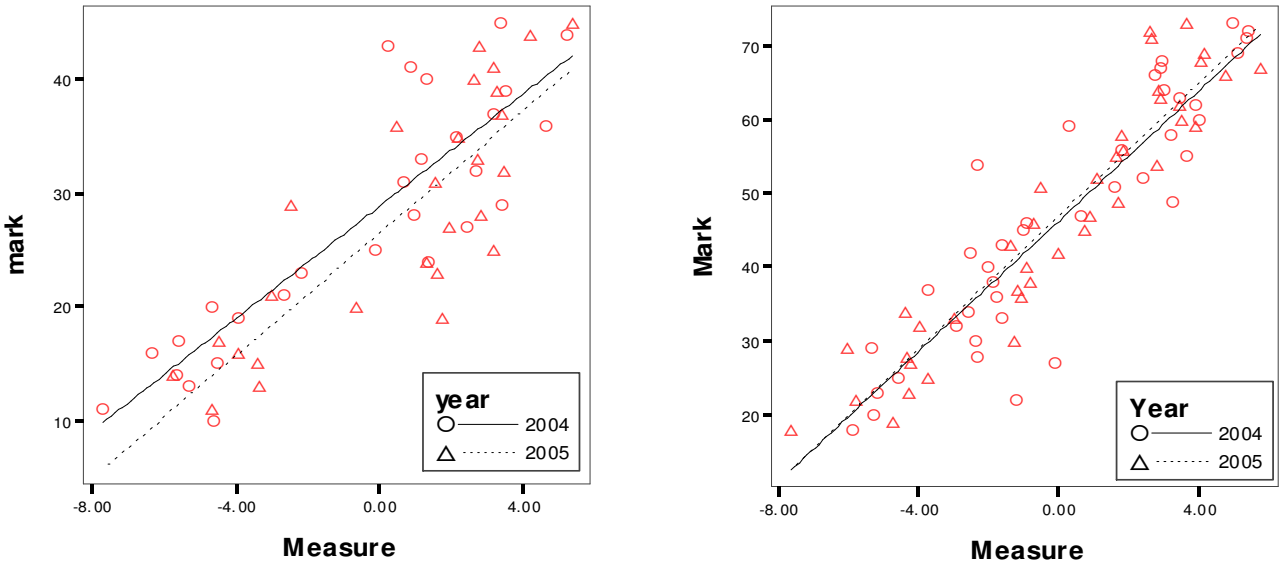
The script and judge measures from the final FACETS runs for both higher and foundation tier are detailed in Appendix B.

Mark vs. measure

There was generally good agreement between the mark and the measure in both years and both tiers. The correlation coefficients for the foundation tier scripts were 0.87 in 2004 and 0.83 in 2005. For the higher tier they were 0.90 in 2004 and 0.94 in 2005. The slightly larger correlations for the higher tier may be a function of it having a larger mark range than that of the foundation tier.

There was one foundation tier script from 2005 (number 37) which lost all of its comparisons and therefore a measure for it could not be estimated. FACETS automatically excludes all judgements involving such a script.

Putting the two years' data onto one graph and adding regression lines allows comparisons between years. This is shown in Figure 2.



Foundation Tier Higher Tier

Figure 2: Comparison of 2004 and 2005 scripts.

For the foundation tier there is a clear difference between the two years with the 2004 regression line higher than the 2005 line, particularly at lower marks. This suggests that a performance of equivalent standard received a higher mark in 2004 than in 2005. There is very little difference between the two years on higher tier with the 2005 regression line very slightly higher than the 2004 line at higher marks. The regression equations for the lines of best fit in figure 2 are shown in Table 1.

Table 1: Regression lines predicting mark from measure.

	Y	Equation	R^2
Foundation Tier	2004 Mark	28.85 + 2.47 x Measure	0.76
	2005 Mark	26.47 + 2.70 x Measure	0.69
Higher Tier	2004 Mark	46.24 + 4.42 x Measure	0.81
	2005 Mark	46.95 + 4.50 x Measure	0.88

By reading off the graph in figure 2 or by inserting values into the regression equations in table 1 it is possible to determine pairs of marks corresponding to equivalent performance in the different years. Thus, a performance gaining 30 marks on the foundation tier in 2004 equates to a mark of 28 in 2005. Such a disparity could be explained in terms of greater accessibility of questions or greater leniency of marking in the 2004 session. Discussion of this disparity in relation to the awarding decisions is in a later section. For the higher tier the marks are very nearly identical over most of the mark range, although there is a small difference at higher marks. Thus, a mark of 60 in 2004 equates to a mark of 61 in 2005.

Judge agreement

It was also possible to investigate the extent to which judges were in agreement with one another. Table 2 shows the number of individual judgements with a z-statistic greater than 2.5.[3]

[3] z-statistic is the standardised residual i.e. residual standardised by its standard error. It is expected to approximate to a unit normal distribution. The residuals are obtained by comparing the observed value of each paired comparison with the expectation derived from the model.

These indicate the extent to which judges disagreed with other judges as a group, and are part of the analysis of the model fit undertaken by FACETS.

Table 2: Misfitting Judgements by Judge

	Foundation	Higher	*% of total judgements*
Judge	2004/5	2004/5	
1	0	4	*0.8%*
2	2	2	*0.8%*
3	1	9	*2.0%*
4	0	1	*0.2%*
5	0	3	*0.6%*
6	7	2	*1.8%*
7	4	7	*2.2%*
Total	*14*	*28*	
Average per pack	*0.40*	*0.67*	

Individually, judges 3, 6 and 7 had the highest number of misfitting judgements. Overall, the higher tier task had more than foundation (and more per pack), but these were a very small proportion of the total judgements made by each judge[4]. Thus, there was a high degree of agreement between judges.

Comparison of rank order outcomes with award meeting outcomes

How well does the outcome of the rank-ordering of scripts from successive years match the outcome of the award meetings in 2005? The current practice at award meetings is to use expert judgement and statistics to determine the key grade boundaries, and then the other boundaries are determined arithmetically by linear interpolation. For the foundation tier, the key boundaries are F and C and at higher tier they are A, C and D. Using the regression equations above, we can map the standard from one year to another. In other words, taking the 2004 mark corresponding to each of the key grade boundaries, the equivalent mark in 2005 (in terms of judged standard of performance) can be determined.

We must stress at this point that we are not suggesting that a difference in the outcomes from an award meeting and from rank ordering implies a lack of validity in either outcome. They are two different tasks that draw on different sources of information. An award meeting (as stipulated by the QCA Code of Practice) involves judgments about the un-cleaned scripts (i.e. including their marks) and several sources of statistical information, whereas the rank ordering is solely a judgemental comparison of scripts.

Table 5: Foundation tier grade boundaries from 2004 and 2005 awarding decisions for unit 2431 compared with rank order outcomes.

		F boundary	C boundary
2004 boundaries	Awarding meeting	20	40
2005 boundaries	Awarding meeting	21	40
	Rank order	17	38

[4] As a percentage of the total number of judgements made in each pack the numbers are very low. E.g. 0.67/45 = 1.5%. In a Normal distribution 1.24% of the observations have $z \geq 2.5$

Table 5 shows a discrepancy between the rank order outcome and the award meeting, at both the C and F grade boundaries. The rank ordering implies that the paper was easier or the marking less stringent in 2004, particularly at the lower end, as a script of equivalent perceived quality achieved a higher mark than in 2005. However, this was not the view of the award meeting as the C boundary remained the same and the F boundary was *increased* by one mark in 2005.

Table 6: Higher tier grade boundaries from 2004 and 2005 awarding decisions for unit 2431 compared with rank order outcomes.

		D boundary	C boundary	A boundary
2004 boundaries	Awarding meeting	22	33	56
2005 boundaries	Awarding meeting	26	36	56
	Rank order	22	33	57

As with the foundation tier, Table 6 shows some differences between the award meeting boundaries and those suggested by rank ordering. At D and C the rank order suggests boundaries at marks well below the actual ones, although there is fairly good agreement at the A boundary. In other words the rank ordering implied the difficulty of the two papers was equivalent in both years, whereas the award meeting concluded that the 2004 paper was harder and/or more stringently marked at the lower end (but equivalent at A).

Discussion

The primary purpose of this research was to investigate the use of a rank-ordering method of standard maintaining on a paper with long answer essay type questions, as opposed to the short answer question papers used in previous research (Bramley, 2005; Black & Bramley, 2006). The evidence is that the method worked well. It produced good correlations between the measures and the original mark for both foundation and higher tier. The data fit the Rasch model well, with only a few misfitting judgements. However, there was no evidence that the rank ordering method worked better with the essay type questions, as hypothesised in the introduction to this paper – the correlations with original mark and misfitting judgements were comparable with the results from previous research.

We also looked at how the outcomes from the rank-ordering exercise compared to the outcomes from the 2005 awarding meeting. In predicting the grade boundaries for 2005, there was discrepancy at the C and D boundary on foundation tier and the C and F boundaries on higher tier compared to the award meeting outcome. However, the differences were really quite small, between 1 mark and 4 marks. It is worth noting that the two occasions where the difference between the two methods was greatest were on the lowest boundaries, D on higher and F on foundation. The larger difference here may be a consequence of the more uneven and idiosyncratic performance of candidates at lower marks making the scripts harder to judge.

It is argued that these differences will inevitably occur due to the different nature of the task and the different information which feeds into the decision; in particular, the use of statistics at the awarding meeting to direct the decision. This seems to be the crux of standard maintaining using expert judgement; because it is at heart a matter of subjective judgement, and because the difference in quality between scripts a few marks apart is likely to be small, the outcome may be different on different occasions. These differences may be small in terms of marks, but could

have a substantial effect on the percentages achieving each grade which, as Cresswell (2000) argues, is highly unlikely to occur in exams with large entries between one year and the next. Hence, the use of expert judgement alone (as in rank ordering) is perhaps not viable. Having said that, rank ordering still has a potential role in making awarding decisions, alongside the statistics. For a fuller discussion on the use of rank ordering in awarding see Black & Bramley (2006). Briefly, it has several advantages over the expert judgement method currently used;

- It is a purer form of expert judgement, comparing one script with another rather than with an internalized standard.
- The leniency and severity of different judges is eliminated.
- Social dynamics would not influence the decisions.
- Judges look at scripts over most of the mark range, not just at key boundaries.
- Misfitting scripts or judges could be identified and if necessary removed.

However, since rank ordering is a relatively new technique applied in this context there are a few methodological issues yet to be resolved – these are discussed below.

Strictly speaking the conversion of ranked data to paired comparisons violates the requirement that paired comparisons be independent, since if script A beats B and script B beats C then under rank ordering we already know the outcome of the paired comparison of A and C (i.e. it is not independent). By assuming that the comparisons *are* independent we under-estimate the standard error of the measures and thus inflate the separation reliability (see Linacre, 2006). However, since our estimates of separation reliability are so high, this is unlikely to be a serious problem. Bramley (2005) found little substantive difference between the measures estimated by analysing rank-ordered data as paired comparisons compared with analysing them as partial credit scales.

The instructions to judges and their interpretation of these instructions are very important. Whether or not judges take account of the difficulty of the two different question papers when making judgments about perceived quality of performance will affect the outcome. Some researchers have expressed doubts as to whether even expert judges are capable of making this allowance:
"It is, after all, the central conundrum of testing educational attainment: the relative merits of easier tasks done well and harder tasks done moderately." (Adams & Pinot de Moira, 2000).

Since judges are allowed to choose their own method for creating their rankings, individual differences in strategy choice may have an effect on the outcome. The judges in this exercise were asked for informal feedback at the end of the process in terms of how easy they found the task and what their strategy was. Interestingly, of the three responses received, two had very similar strategies. Both these judges gave each script a 'grade' as they went along, and a '+' or '-'. This would immediately reduce the cognitive load involved as the scripts are essentially grouped and comparisons made within those groups. One of the judges went further by giving some scripts '++' or '--'. By giving the scripts a 'grade' in this way the judges are also implicitly taking account of the difficulty of the paper. However, this strategy is actually making use of an internalised abstract standard, thus thwarting one of the main purposes of using rank-ordering!

It is still unclear if there is an ideal number of judges for a rank ordering exercise. Bramley (2005) used twelve judges, Black & Bramley (2006) used nine, and this research used seven. Each of these has produced a seemingly valid set of measures with reasonable correlation with mark. Obviously the greater the number of comparisons per script, the more accurately its measure can be estimated, but this needs to be traded off against the costs of increasing the number of judges or the number of judgments required from each judge.

The number of scripts per pack is another issue that is as yet unresolved. The task demand of holding an idea of the quality of ten scripts in mind at once is of concern. However, reducing the

number of scripts per pack would substantially reduce the number of paired comparisons. Eight scripts would generate 28 comparisons, six scripts would generate only 15. This reduction in data would have to be compensated by having more packs and/or judges.

There is also the question of the mark range within the packs. This research had packs with ranges slightly wider on average than Black & Bramley (2006). However, the overall fit of the model seems not to have been improved by wider pack ranges, with similar correlations and misfitting judgements to Black & Bramley (*op. cit.*). Hence it may be that the effect of pack range is minimal, within certain limits at least.

Further research at Cambridge Assessment is currently investigating the effect of varying these 'design parameters' of a rank-ordering exercise on the stability of the outcome.

References

Adams, R. & Pinot de Moira, A. (2000). *A Comparability Study in GCSE French including parts of the Scottish Standard Grade Examination. A study based on the Summer 1999 examination.* Organised by WJEC and AQA on behalf of the Joint Forum for GCSE and GCE.

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

Bell, J.F., Bramley, T., and Raikes, N. (1998) Investigating A-level mathematics standards over time. *British Journal of Curriculum and Assessment*, 8, 7-11.

Black, B. & Bramley, T. (2006) *An investigation and cross-validation of 2004 and 2005 standard setting in GCE A-level Psychology using a rank-ordering method,* Paper presented at the British Educational Research Association annual conference, University of Warwick, September 2006.

Bramley, T. (2005) A Rank-Ordering Method for Equating Tests by Expert Judgement. *Journal of Applied Measurement,* 6, 202-223.

Cresswell, M.J. (2000). The role of public examinations in defining and monitoring standards. In H. Goldstein & A. Heath (Eds) *Educational Standards.* (Oxford, Oxford University Press) 69-104.

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

Greatorex, J. (2003). *What happened to limen referencing? An exploration of how the Awarding of public examinations has been and might be conceptualized.* Paper presented at the British Educational Research Association annual conference, Heriot-Watt University, September 2003.

Jones, B. & Meadows, M. (2004). *Report of the inter-Awarding Body comparability study into GCSE Religious Studies (full course), Summer 2003.* A study sponsored and undertaken by the AQA with support and advice from the WJEC. DRAFT, July 2004.

Linacre, J.M. (2005). *Facets Rasch measurement computer program*. (Chicago, Winsteps.com).

Linacre, J.M. (2006). Rasch Analysis of Rank-Ordered Data. *Journal of Applied Measurement,* 7(1), 129-139.

Murphy, R J L., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmut, J., and Gower, R. (1996). *The Dynamics of GCSE Awarding: Report of a project conducted for the School Curriculum and Assessment Authority.* (London, SCAA).

Pollitt, A. and Crisp, V. (2004). *Could comparative judgements of script quality replace traditional marking and improve the validity of exam questions?* Paper presented at the British Educational Research Association annual conference, UMIST, Manchester, September 2004.

Qualifications and Curriculum Authority (2006). *GCSE, GCE, VCE, GNVQ and AEA Code of Practice 2006/7* (London, QCA).

Qualifications and Curriculum Authority (2006), QCA's Review of Standards, Available online at: http://www.qca.org.uk/downloads/qca-06-2374_QCAs-review-of-standards.pdf (accessed 27 July 2006).

Scharaschkin, A. & Baird, J-A. (2000) The Effects of Consistency of Performance on A Level Examiners' Judgements of Standards*, British Educational Research Journal*, 26, 333-357.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review, 3*, 273-286.

Thurstone, L. L. (1931). Rank order as a psychophysical method. *Journal of Experimental Psychology, 14,* 187-201. Chapter 10 in Thurstone, L.L. (1959). *The measurement of values.* University of Chicago Press, Chicago, Illinois.

Wright, B. & Stone, S. (1979) *Best Test Design: Rasch Measurement* (Mesa Press, Chicago).

Appendix A – Pack design

Pack Design 2431 Foundation Tier

	2004 Boundaries				G					F					E					D					C				
	2005 Boundaries					G					F				E					D					C				
	Mark Points	10	11	13	14	15	16	17	19	20	21	23	24	25	27	28	29	31	32	33	35	36	37	39	40	41	43	44	45
Pack	Judge																												
1	5																				45		45			45		45	45
1	2																				5	4	5	4	45	5	4	5	4
1	6																			4	4	4	4	45	5	5	5		5
1	1																		4		5	4	45	5	4	5	5	4	
1	3																		5	5	5	5	5	4	4	4	4	4	
1	7																		5	4	4	45	5	5	4	4	5		
1	4																		4	5	4	45	5	4	5	5	4		
2	6														4	5	4	5	45	5	4	5	4						
2	1													4	4	4	4	45	5		5	5	5						
2	5													5	4	4	5	5	4	45	4	5							
2	7													4	5	5	4	4	45	5	5		4						
2	4												5	5	5	5	5	4	4	4	4		4						
2	3												4	5	5	4	4	5	4	5	45								
2	2												45	4	4	5	5	45	5	4									
3	3									45	4		5	4	5	4	5	45											
3	7								4	5	4	5	45	5	4	5	4												
3	5							5	5	5	5	5	4	4	4	4		4											
3	6								4	45	4	5	5	5	4	4		5											
3	1							45	5	4	5	4	45	4		5													
3	2							4	4	4	4	4	5	5	5	5	5												
3	4							5	4	45	4	5	4	5	5	4													
4	3			5	4	5	4	45	5		45	4																	
4	7		4	4	4	4	4	5	5	5	5	5																	
4	4	4	5	4	45		4	5		5	45																		
4	6		4	5	5	4	45	4	5	5	4																		
4	1	5	5	5	5	5	4	4	4	4		4																	
4	2	4	5	4	5	4	5	4	5	5	4																		
4	5	5		4	4	5	45	45		4	5																		
5	1			4		4	5				4	5			5					5	4			4	5				
5	2				4	5	4				5					4				5	4			5		4	5		
5	3	4	5				5		4			5			4	5				4			4	5					
5	4							4	5					4			4	5			5			5	4			4	5
5	5		5		5				4			4	5				4		5			4			45				
5	6							5				4		4	5			4			5		5		4			5	4
5	7	4		5			5	4				4		5				5	4					4	5				

Pack Design 2431 Higher Tier

	2004 Boundaries				D								C						B							A					A*				
	2005 Boundaries						D								C						B							A				A*				
	Mark Points	18	19	20	22	23	25	27	28	29	30	32	33	34	36	37	38	40	42	43	45	46	47	49	51	52	54	55	56	58	59	60	62	63	64	66	67
Pack	Judge																																				
6	4																																5	5	5	5	45
6	5																																5	4	5	45	4
6	1																															5		4	5	4	5
6	7																														4		4	4	4	4	5
6	6																														5	4		45		5	4
6	2																														45	5	4		5	4	5
6	3																															45	4	45	45		5
7	7																											45		45	4	5	4	5	4	5	
7	6																									5	4	4	4	5	4	5		5	5	4	
7	4																										5	4	5	4	5	4	4	5		45	
7	1																									5	5	5	5	4	4	4	5	4	4		
7	5																									45	4	5	45		5	4	5	4			
7	3																									45	4	5	4	5		5	4	45			
7	2																								4	4		4		45	45	5	5	5			
8	6																					4	5	4	5	5	4	5	45	4							
8	2																					5	5	5	5	5	4	4	4	4	4						
8	1																					5	4	4	5	4	5	4	5		45						
8	4																				45	4	5	5	4	5	4	5		4							
8	3																				45	5	4	45		4	5	5	4								
8	5																			4		4	4	45		4	5	5	5	5							
8	7																			5	4	45	5	4	5	45	4										
9	2																	45	4	4	5	4	5	4	5	5											
9	5															4	4	5	5	4	5	5	4	5	4												
9	1															4	4	4	4	4	5	5	5	5	5												
9	7															45	4	5	4	5	4		5	45													
9	4														4	5			45	45		5	45		4												
9	3															5	5	5	5	5	4	4	4	4	4												
9	6														45		5	4	5	4	5	45	4														
10	1											4	5	4	45	5			4	5	45																
10	6										4		45	5	4	5	5	4	5	4																	
10	5									4		4	4	4	4	5	5	5	5	5																	
10	3									4	5	4	5	4	5	4	5	5	4																		
10	4									5		5	5	5	5	4	4	4	4	4																	
10	7									4	5	45		45		4		45	5																		
10	2									5		45	4	5	4	5	45	4																			
11	3				4		4	5	4	5	5	4	45	5																							
11	4		5			5	5	5	45	4	4	4		4																							
11	7			4	5	5	4			45	4	45	5																								
11	6	4	5			4	5	4	5	5	4	5	4																								
11	1	4		4	4	4	45	5	5		5	5																									
11	5	5		4	5	4	5	5	4	45		4																									
11	2	45	5		4	5		4	5	4	45																										

Appendix B – FACETS output, Foundation Tier

GCSE English Thurstone rankings 06-08-2006 11:59:35
Table 7.1.1 Judge Measurement Report (arranged by mN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	N	Judge
113	225	.5	.50	.00	.20	1.33	3.0	1.09	.7	.55	7	AC
108	216	.5	.50	.00	.19	.93	-.7	.65	.4	1.12	5	CB
108	216	.5	.50	.00	.20	.75	-2.7	.45	.2	1.38	1	RC
104	208	.5	.50	.00	.20	.90	-1.0	.64	.2	1.17	3	BD
104	208	.5	.50	.00	.20	1.19	1.9	1.81	.9	.64	6	JR
113	225	.5	.50	.00	.20	.84	-1.6	.64	.6	1.20	2	CM
113	225	.5	.50	.00	.19	1.03	.3	.67	.2	1.01	4	KW
108.8	217.6	.5	.50	.00	.20	1.00	-.1	.85	.5		Mean (Count: 7)	
3.6	7.1	.0	.00	.00	.00	.19	1.9	.43	.3		S.D.	

RMSE (Model) .20 Adj S.D. .00 Separation .00 Separation Reliability 44E4
Fixed (all same) chi-square: .0 d.f.: 6 significance (probability): 1.00

GCSE English Thurstone rankings 06-08-2006 11:59:35
Table 7.3.1 Script Measurement Report (arranged by mN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Nu	Script
14	27	.5	1.00	5.74	.66	1.15	.4	.76	.9	.94	53	S5_4413an (mark 45)
18	36	.5	1.00	5.60	.58	.71	-.6	.42	.5	1.23	28	S4_9485st (mark 44)
27	54	.5	.99	5.00	.41	1.08	.4	.82	1.4	.95	4	S4_3053ad (mark 36)
14	27	.5	.99	4.57	.53	1.08	.3	.91	2.1	.91	47	S5_1120jd (mark 44)
27	54	.5	.98	3.87	.36	1.16	.9	.92	1.8	.72	23	S4_9187ap (mark 39)
32	63	.5	.98	3.83	.34	1.03	.2	.89	1.0	.96	49	S5_3193bm (mark 32)
32	63	.5	.98	3.83	.34	1.07	.4	.88	1.4	.91	25	S4_9194mp (mark 29)
36	72	.5	.98	3.78	.30	1.02	.1	.85	.4	1.02	39	S5_0190ps (mark 37)
14	27	.5	.98	3.72	.48	.68	-1.5	.49	1.2	1.61	27	S4_9347ap (mark 45)
26	52	.5	.97	3.59	.37	1.00	.0	.80	1.9	1.01	44	S5_0478mi (mark 39)
36	72	.5	.97	3.57	.35	1.02	.1	.70	1.4	.97	45	S5_0502jw (mark 25)
23	45	.5	.97	3.56	.34	.97	-.1	.92	-.2	1.09	42	S5_0266hj (mark 41)
31	61	.5	.97	3.54	.31	.88	-1.0	1.09	.8	1.21	8	S4_5084kh (mark 37)
35	70	.5	.96	3.17	.33	.92	-.5	.60	1.1	1.20	48	S5_2128nm (mark 28)
18	36	.5	.96	3.13	.41	.78	-1.3	.65	1.8	1.47	52	S5_4063hm (mark 43)
32	63	.5	.96	3.11	.33	.92	-.4	.70	.5	1.16	15	S4_9055ef (mark 27)
36	72	.5	.96	3.10	.29	1.01	.1	.81	1.0	1.02	55	S5_5119ln (mark 33)
36	72	.5	.96	3.07	.28	1.00	.0	.92	.6	.99	17	S4_9093gs (mark 32)
27	54	.5	.95	3.02	.36	1.26	1.6	1.14	1.5	.34	40	S5_0243dp (mark 40)
27	54	.5	.93	2.64	.55	1.24	.9	.88	2.4	.61	41	S5_0250rl (mark 19)
41	81	.5	.93	2.56	.27	.95	-.3	.87	.3	1.09	50	S5_3275ch (mark 35)
45	90	.5	.93	2.53	.26	1.09	.8	.94	.6	.83	24	S4_9193sp (mark 35)
36	72	.5	.91	2.30	.32	1.00	.0	.76	.6	.99	30	S5_0004jr (mark 27)
27	54	.5	.88	2.03	.42	1.44	2.1	4.00	1.9	-.13	33	S5_0021ra (mark 23)
41	81	.5	.87	1.93	.29	1.10	.7	.90	.4	.85	35	S5_0049sd (mark 31)
27	54	.5	.85	1.74	.37	.93	-.3	.67	.5	1.14	6	S4_4295dp (mark 24)
32	63	.5	.84	1.67	.37	.76	-1.1	.62	.2	1.25	22	S4_9166ss (mark 40)
36	72	.5	.84	1.66	.35	1.02	.1	.69	.4	.99	56	S5_6095rl (mark 24)
26	52	.5	.80	1.40	.39	.98	.0	.74	.0	1.07	9	S4_7030gd (mark 33)
36	72	.5	.79	1.33	.33	1.10	.6	.82	.4	.89	20	S4_9153jp (mark 28)
18	36	.5	.77	1.21	.51	.93	-.1	.54	.5	1.14	19	S4_9105mp (mark 41)
32	63	.5	.73	1.01	.35	.79	-1.1	.57	-.3	1.28	14	S4_9050nc (mark 31)
27	54	.5	.70	.86	.40	.86	-.5	.62	-.7	1.17	51	S5_3386jn (mark 36)
14	27	.5	.65	.62	.67	.97	.0	.67	.0	1.06	21	S4_9157dn (mark 43)
36	72	.5	.54	.17	.40	.86	-.4	.86	.1	1.09	5	S4_4146as (mark 25)
40	79	.5	.30	-.85	.56	.93	.0	.81	.2	1.01	46	S5_0646ht (mark 20)
23	45	.5	.06	-2.68	.80	.41	-1.3	.07	-.2	1.33	34	S5_0048sf (mark 29)
31	62	.5	.06	-2.70	.54	.93	.0	.52	.2	1.07	11	S4_9005pa (mark 23)
44	88	.5	.04	-3.09	.43	1.19	.7	2.38	1.2	.81	2	S4_0368jw (mark 21)
31	62	.5	.02	-3.73	.47	.95	.0	.58	.2	1.06	29	S5_0001mc (mark 21)
17	33	.5	.02	-3.82	.54	1.32	1.0	1.03	.8	.68	36	S5_0075sg (mark 13)
17	34	.5	.02	-4.13	.51	.71	-1.0	.43	.5	1.37	43	S5_0411ls (mark 15)
30	60	.5	.01	-4.42	.48	1.02	.1	.39	1.1	1.04	12	S4_9007cb (mark 19)
21	42	.5	.01	-5.18	.45	.91	-.4	.51	1.9	1.21	54	S5_5101kj (mark 16)
17	34	.5	.01	-5.22	.52	1.65	2.2	1.73	1.7	-.08	10	S4_8057rh (mark 15)
17	34	.5	.01	-5.29	.49	1.06	.3	.76	2.1	.85	7	S4_4526lp (mark 10)
21	42	.5	.00	-5.34	.43	.75	-1.3	.45	2.1	1.51	38	S5_0131mm (mark 11)
31	61	.5	.00	-5.42	.44	.76	-1.0	.32	1.1	1.34	26	S4_9195rt (mark 20)
36	71	.5	.00	-5.69	.39	1.36	1.7	.89	1.6	.49	32	S5_0021ab (mark 17)
22	44	.5	.00	-6.13	.41	.89	-.6	.54	1.9	1.29	16	S4_9080am (mark 13)
39	77	.5	.00	-6.33	.37	.92	-.3	.41	2.0	1.18	13	S4_9049sb (mark 17)
21	42	.5	.00	-6.45	.45	.97	.0	.78	2.4	1.02	31	S5_0015ec (mark 14)
22	44	.5	.00	-6.46	.44	1.00	.0	.81	2.2	.97	1	S4_0303ja (mark 14)
30	59	.5	.00	-7.09	.41	.87	-.5	.49	2.0	1.18	3	S4_1068jh (mark 16)
8	16	.5	.00	-8.45	1.08	1.23	.5	1.66	4.5	.70	18	S4_9098jl (mark 11)
9	18			(-9.98	1.85)	Minimum					37	S5_0098sm (mark 10)
27.4	54.7	.5	.57	.00	.43	.99	.0	.84	1.1		Mean (Count: 56)	
8.9	17.9	.0	.43	4.01	.14	.20	.8	.55	.9		S.D.	

RMSE (Model) .45 Adj S.D. 3.99 Separation 8.76 Separation Reliability .99
Fixed (all same) chi-square: 4666.5 d.f.: 54 significance (probability): .00
Random (normal) chi-square: 53.8 d.f.: 53 significance (probability): .44

GCSE English Higher Tier Thurstone rankings 06-08-2006 14:14:49
Table 7.1.1 Judge Measurement Report (arranged by mN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	N	Judge
135	270	.5	.50	.00	.16	1.07	.9	1.22	.9	.88	1	RC
135	270	.5	.50	.00	.16	1.18	2.2	1.84	2.3	.62	3	BD
131	261	.5	.50	.00	.17	.88	-1.4	.73	-.7	1.18	2	CM
135	270	.5	.50	.00	.16	1.06	.7	1.13	.5	.88	7	AC
135	270	.5	.50	.00	.16	.85	-2.1	.69	-1.6	1.25	4	KW
135	270	.5	.50	.00	.16	.83	-2.2	.66	-.9	1.26	6	JR
118	236	.5	.50	.00	.17	1.00	.0	.89	-.3	1.01	5	CB
131.9	263.9	.5	.50	.00	.16	.98	-.3	1.02	.0		Mean (Count: 7)	
5.9	11.8	.0	.00	.00	.00	.12	1.6	.39	1.3		S.D.	

RMSE (Model) .16 Adj S.D. .00 Separation .00 Separation Reliability 26E4
Fixed (all same) chi-square: .0 d.f.: 6 significance (probability): 1.00

GCSE English Higher Tier Thurstone rankings 06-08-2006 14:14:49
Table 7.3.1 Script Measurement Report (arranged by mN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Nu	Script
14	27	.5	1.00	5.79	.53	.84	-.4	.58	-.7	1.23	72	S5_2266dl (mark 67)
18	36	.5	1.00	5.44	.44	.91	-.3	.74	-.6	1.16	26	S4_9092ms (mark 72)
18	36	.5	1.00	5.41	.44	1.16	.7	1.27	.7	.79	31	S4_9159cw (mark 71)
18	36	.5	.99	5.15	.41	1.04	.2	.86	-.3	1.00	22	S4_9005pa (mark 69)
9	18	.5	.99	4.99	.56	.90	-.3	.93	.0	1.16	9	S4_1079wc (mark 73)
32	63	.5	.99	4.78	.31	.98	-.1	.94	-.1	1.04	58	S5_0310sb (mark 66)
14	27	.5	.98	4.17	.44	1.33	1.5	1.39	1.5	.33	62	S5_0457rt (mark 69)
23	45	.5	.98	4.08	.34	1.10	.7	1.11	.7	.76	74	S5_3186sf (mark 68)
23	45	.5	.98	4.03	.35	.94	-.3	.90	.0	1.13	19	S4_4287dm (mark 60)
23	45	.5	.98	3.93	.35	.83	-1.1	.72	-.3	1.37	6	S4_1011ac (mark 62)
18	36	.5	.98	3.92	.42	1.28	1.4	1.24	.5	.50	50	S5_0121kl (mark 59)
27	54	.5	.98	3.69	.35	1.06	.4	1.30	.6	.83	4	S4_0425pk (mark 55)
9	18	.5	.98	3.67	.52	.95	-.1	.92	-.2	1.15	66	S5_1013kb (mark 73)
27	54	.5	.97	3.55	.32	.82	-1.3	.71	-.4	1.40	77	S5_4340am (mark 60)
27	54	.5	.97	3.48	.31	.90	-.8	.79	-.2	1.29	27	S4_9099dn (mark 63)
27	54	.5	.97	3.48	.32	1.25	1.8	1.35	.9	.38	76	S5_4064sg (mark 62)
36	72	.5	.96	3.30	.42	.97	.0	.65	-.1	1.03	17	S4_2422rp (mark 49)
32	63	.5	.96	3.24	.32	1.09	.6	.90	.0	.90	18	S4_3019rf (mark 58)
32	63	.5	.95	3.01	.29	.96	-.3	.95	-.1	1.10	35	S4_9209ej (mark 64)
14	27	.5	.95	2.98	.49	.88	-.4	.74	-.5	1.21	14	S4_2319fb (mark 68)
32	63	.5	.95	2.96	.30	.83	-1.3	.79	-.4	1.32	65	S5_0633rw (mark 63)
18	36	.5	.95	2.92	.39	.99	.0	.92	-.1	1.04	23	S4_9025se (mark 67)
18	36	.5	.95	2.87	.41	.87	-.6	.67	-.6	1.29	79	S5_7012jm (mark 64)
23	45	.5	.94	2.83	.37	1.14	.8	1.00	.1	.80	41	S5_0001ja (mark 54)
18	36	.5	.94	2.80	.38	1.12	.8	1.30	.8	.59	1	S4_0003pa (mark 66)
18	36	.5	.94	2.69	.40	.96	-.2	.81	-.5	1.14	46	S5_0085ig (mark 71)
14	27	.5	.93	2.63	.46	.88	-.5	.75	-.6	1.25	70	S5_2017ck (mark 72)
32	63	.5	.92	2.46	.32	.84	-1.0	.66	-.7	1.29	20	S4_7468tg (mark 52)
27	54	.5	.87	1.91	.34	.80	-1.1	.71	-.8	1.29	55	S5_0166kp (mark 56)
23	45	.5	.86	1.83	.37	.84	-.8	.70	-.8	1.29	43	S5_0071mm (mark 58)
27	54	.5	.86	1.83	.34	1.15	.9	2.15	2.4	.58	21	S4_9001ja (mark 56)
32	63	.5	.85	1.73	.35	1.34	1.8	2.11	1.6	.39	67	S5_1046eh (mark 49)
36	72	.5	.84	1.68	.29	.98	-.1	.98	.0	1.02	49	S5_0121er (mark 55)
23	45	.5	.84	1.62	.42	1.39	1.5	1.32	.7	.61	8	S4_1059tb (mark 51)
41	81	.5	.75	1.11	.30	.75	-1.6	.70	-1.0	1.28	51	S5_0124rr (mark 52)
36	72	.5	.72	.95	.33	1.10	.6	1.09	.3	.88	73	S5_3077st (mark 47)
32	63	.5	.69	.79	.35	.95	-.2	.89	-.1	1.05	45	S5_00841e (mark 45)
32	63	.5	.66	.69	.33	.78	-1.1	.60	-1.4	1.29	13	S4_2313hw (mark 47)
27	54	.5	.58	.31	.44	.89	-.3	.95	.0	1.08	40	S4_9490ms (mark 59)
9	17	.5	.51	.05	1.14	.69	-.2	.16	1.1	1.28	33	S4_9170ss (mark 27)
31	61	.5	.51	.04	.33	1.02	.1	.98	.1	.96	60	S5_0360nf (mark 42)
27	54	.5	.38	-.48	.38	.83	-.7	.58	-.8	1.24	48	S5_0093rn (mark 51)
18	36	.5	.36	-.56	.51	1.60	2.0	2.60	1.2	.07	24	S4_9075ce (mark 33)
36	72	.5	.35	-.64	.31	1.00	.0	.85	-.2	1.02	71	S5_2217lg (mark 46)
26	52	.5	.32	-.73	.34	.75	-1.8	.64	-.4	1.52	47	S5_0091ag (mark 38)
32	63	.5	.30	-.86	.34	1.11	.6	.95	.0	.88	37	S4_9288dh (mark 46)
31	61	.5	.29	-.89	.32	1.06	.4	1.15	.4	.84	78	S5_5057bs (mark 40)
27	54	.5	.28	-.95	.36	1.05	.3	1.00	.1	.92	11	S4_2182bd (mark 45)
18	36	.5	.27	-1.00	.41	.78	-1.1	.62	-.2	1.41	44	S5_0080mh (mark 36)
31	61	.5	.25	-1.10	.31	.92	-.5	.97	.0	1.13	59	S5_0331na (mark 37)
13	26	.5	.23	-1.19	.70	.95	.0	.38	.0	1.13	39	S4_9329jp (mark 22)
22	44	.5	.23	-1.22	.46	1.26	1.1	2.69	1.4	.47	80	S5_7996et (mark 30)
26	52	.5	.21	-1.35	.34	.84	-1.1	1.01	.1	1.24	52	S5_0136jp (mark 43)
36	72	.5	.17	-1.57	.31	.81	-1.3	.67	-.6	1.29	2	S4_0037tb (mark 43)
26	52	.5	.15	-1.72	.33	1.23	1.6	1.29	.8	.48	10	S4_2113dp (mark 36)
23	45	.5	.14	-1.81	.39	.79	-1.1	1.19	.4	1.22	32	S4_9166lm (mark 38)
32	63	.5	.13	-1.93	.32	.74	-1.8	.56	-.9	1.43	12	S4_2274gh (mark 40)
32	63	.5	.09	-2.27	.62	1.02	.1	1.13	.5	.97	5	S4_0897ch (mark 54)
35	70	.5	.08	-2.40	.32	1.05	.3	1.51	1.1	.88	3	S4_0183em (mark 32)
32	63	.5	.08	-2.47	.34	1.18	1.0	1.07	.3	.76	30	S4_9119sd (mark 42)
22	44	.5	.08	-2.47	.47	1.04	.2	.64	-.3	1.04	16	S4_2405cm (mark 30)
13	25	.5	.07	-2.57	.62	1.20	.6	.86	.1	.87	28	S4_9099sh (mark 28)
22	43	.5	.06	-2.69	.39	1.03	.1	1.64	1.5	.89	29	S4_9108lp (mark 34)
27	54	.5	.05	-2.97	.37	1.05	.3	1.36	.8	.89	57	S5_0266ab (mark 33)

27	54	.5	.02	-3.78	.46	.86	-.3	.61	-.1	1.13	38	S4_9297gd (mark 37)
27	54	.5	.01	-4.38	.40	.69	-1.3	.45	-1.3	1.33	54	S5_0157as (mark 32)
17	34	.5	.01	-4.47	.46	.92	-.2	.77	-.3	1.12	75	S5_3349am (mark 25)
23	45	.5	.01	-4.59	.51	.78	-.6	1.52	.8	1.09	61	S5_0451mg (mark 34)
13	26	.5	.01	-4.87	.53	.50	-2.0	.32	-1.1	1.61	63	S5_0496cm (mark 23)
18	35	.5	.01	-5.02	.45	.98	.0	.71	-.1	1.08	56	S5_0242eu (mark 28)
17	34	.5	.01	-5.02	.46	1.04	.2	1.98	1.5	.80	53	S5_0136mc (mark 27)
14	27	.5	.01	-5.22	.53	1.49	1.6	2.93	1.9	.23	25	S4_9086jh (mark 25)
13	26	.5	.00	-5.44	.54	.55	-2.0	.34	-.4	1.65	64	S5_0527km (mark 19)
13	25	.5	.00	-6.30	.53	1.02	.1	.81	.3	.99	7	S4_1045ah (mark 23)
26	51	.5	.00	-6.40	.47	1.50	1.7	3.06	1.5	.23	34	S4_9200wb (mark 29)
13	25	.5	.00	-6.42	.53	.82	-.7	.56	-.2	1.34	36	S4_9215cd (mark 20)
13	26	.5	.00	-6.72	.61	.80	-.5	.48	.4	1.26	15	S4_2380lm (mark 18)
26	52	.5	.00	-7.11	.53	.92	-.1	.39	.1	1.12	69	S5_1179at (mark 29)
8	16	.5	.00	-7.18	.71	1.07	.3	.84	.3	.92	42	S5_0035jp (mark 22)
5	9			(-8.69	1.91)	Minimum					68	S5_1082dl (mark 18)

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Nu	Script
23.1	46.3	.5	.52	.00	.42	.98	-.1	1.02	.1			Mean (Count: 80)
8.1	16.2	.0	.41	3.52	.13	.20	1.0	.56	.8			S.D.

RMSE (Model) .44 Adj S.D. 3.49 Separation 7.94 Separation Reliability .98
Fixed (all same) chi-square: 5038.6 d.f.: 78 significance (probability): .00
Random (normal) chi-square: 77.8 d.f.: 77 significance (probability): .45