CAMBRIDGE ASSESSMENT

How accurate are examiners' judgments of script quality?

An investigation of absolute and relative judgments in two units, one with a wide and one with a narrow 'zone of uncertainty'

Tim Gill and Tom Bramley

Paper presented at the British Educational Research Association annual conference,
Heriot-Watt University, Edinburgh, September 2008

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

Gill.Tim@cambridgeassessment.org.uk

Abstract

The setting of grade boundaries on components of an assessment at an award meeting can be summarised as follows: a range of scripts on marks around where the key boundaries are likely to lie is scrutinised by the awarders. Starting from the top of the range of marks, awarders determine the lowest mark for which there is consensus that the work is worthy of the higher grade. This mark is the top of the 'zone'. Then, working from the bottom up, they determine the highest mark for which there is consensus the work is not worthy of the higher grade. The mark above this is the bottom of the 'zone'. This gives a range of a few marks where there is not consensus either way (the zone), and the grade boundary lies at some point within this range.

The size of the zone would appear to tell us something about the resolving power of the examiners' judgments. For instance, a zone of 72-69 at the A/B boundary means that there is consensus that scripts with a mark of 72 are worthy of an A and scripts with a mark of 68 are not worthy. This suggests that a mark difference of 4 should be enough for the examiners to be able to distinguish between a pair of scripts, at least at this part of the mark scale.

In an award meeting the marks on the scripts (both for individual questions and the paper total) are clearly visible. This research investigated the accuracy of examiners' judgments (their 'resolving power') when the scripts they looked at were cleaned of all marks. Units from two different examinations were used, one with a relatively wide zone (A-level Physics, 2824) and one with a relatively narrow zone (A-level History, 2589), in order to see whether the same implied resolving power would be achieved in the absence of knowledge of the marks.

We asked examiners to look at pairs of scripts (of varying mark differences) and make three types of judgment about each pair:

i) Absolute – which grade each script was worth
ii) Relative – which of the two scripts was better in terms of the overall quality
iii) Confidence – how confident they were about the judgments in i) and ii).

These judgments have some similarities with the judgmental part of award meetings: at an award meeting examiners make holistic, absolute judgments about whether a script on a particular mark point is worthy of a particular grade (having internalised the grade boundary standard by studying 'archive' scripts from past sessions). By inference they also make holistic, relative judgments: a script that is deemed to be worthy of the higher grade must be considered better in terms of overall quality than one that is only worthy of the lower grade.

We compared the outcomes of the paired comparisons with the 'correct' outcomes, as defined by the original marking and grading of the scripts.

To summarise, we found:

- The examiners had difficulty in accurately judging the grade-worthiness of the scripts when the marks were removed. The overall accuracy of the (absolute) judgments of grade-worthiness was higher for History than Physics.
- The overall accuracy of the relative judgments of script quality was higher than that of the absolute judgments, and in contrast with the absolute judgments the Physics judges were more accurate than the History judges.
- In Physics, the accuracy of the relative judgments increased fairly consistently as the difference in marks between the scripts in each pair increased.
- In History there was a surprising, and inexplicable, inconsistent relationship between mark difference and accuracy of relative judgment – the accuracy increased from 44% to 75% going from a 2-mark difference to a 4-mark difference, but then decreased steadily to 60% at a 12-mark difference before rising again to 86% at a 16-mark difference.
- In both History and Physics the accuracy of relative judgments at mark differences close to those implied by the 'zone' were well short of 100%. The relative size of the zone in

comparison to the grade bandwidth suggested that the History judges would be capable of more accurate relative judgments than the Physics judges, but there was no support for this hypothesis in the data – in fact the opposite seemed to be the case.
- Both History and Physics judges were more confident about their relative than their absolute judgments.
- The accuracy of both the absolute and relative judgments was significantly related to confidence in both subjects.

We note that there were several differences between the experimental situation in this research study and the situation in an awarding meeting – specifically, the marks were removed from the scripts, more scripts were judged, there was no external information to guide judgments, and there was no chance for 'peer pressure' to influence judgments.

In these, arguably preferable, conditions (from the point of view of obtaining a 'pure' judgmental outcome), the fact that the judges were more accurate at making relative, rather than absolute judgments raises some doubts as to whether the current awarding procedures make best use of expert judgment. Alternatives such as rank ordering, or paired comparisons, which require relative rather than absolute judgments, may be a better way of using the examiners' skills, particularly if this is done independently of the external statistical information about mark distributions and cohort characteristics.

1. Introduction

The setting of grade boundaries on components of an assessment at an award meeting as laid out in the QCA Code of Practice (QCA, 2006) can be summarised as follows: a range of scripts on marks around where the key boundaries are likely to lie is scrutinised by the awarders. Starting from the top of the range of marks, awarders determine the lowest mark for which there is consensus that the work is worthy of the higher grade. This mark is the top of the 'zone'. Then, working from the bottom up, they determine the highest mark for which there is consensus the work is *not* worthy of the higher grade. The mark above this is the bottom of the 'zone'. This gives a range of a few marks where there is not consensus either way (the zone), and the grade boundary lies at some point within this range. Awarders then use their collective professional judgment to recommend a single mark where the boundary should lie, drawing on a variety of sources of evidence such as mark distributions, entry patterns, prior attainment, performance in previous sessions, forecast grades, and monitoring and comparability research reports.

The judgmental part of this process relies on the idea that examiners have an 'internal standard' of what, for instance, a borderline A grade script should look like, to which they can refer when making their judgments. This internal standard is an abstraction of the relevant qualities of a borderline script, which derives partly from experience, and partly from the consideration of several archive scripts on the grade boundary that takes place before an awarding meeting. The judgmental part of the awarding process also identifies the difference in marks between scripts that are considered to be definitely worthy of the grade and those that are not. Thus at an award meeting examiners make **absolute**[1] judgments about whether a script on a particular mark point is worthy of a particular grade, and by inference they make a **relative** judgment: a script that is deemed to be worthy of the higher grade must be considered better in terms of overall quality than one that is only worthy of the lower grade. The size of the zone would appear to tell us something about the 'resolving power' of the examiners' judgments. For instance, a zone of 72-69 at the A/B boundary means that there is consensus that scripts with a mark of 72 are worthy of an A and scripts with a mark of 68 are not worthy. This suggests that a mark difference of 4 should be enough for the examiners to be able to distinguish between a pair of scripts, at least at this part of the mark scale.

In an award meeting the marks on the scripts (both for individual questions and the paper total) are clearly visible. This research investigated the 'resolving power' of examiners' judgments when the scripts they looked at were cleaned of all marks. Units from two different examinations were used, one with a relatively wide zone and one with a relatively narrow zone, in order to see whether the same implied resolving power would be achieved in the absence of knowledge of the marks.

We asked examiners to look at pairs of scripts (of varying mark differences) and make three types of judgment about each pair:

i) Absolute – which grade each script was worth
ii) Relative – which of the two scripts was better in terms of the overall quality
iii) Confidence – how confident they were about the judgments in i) and ii).

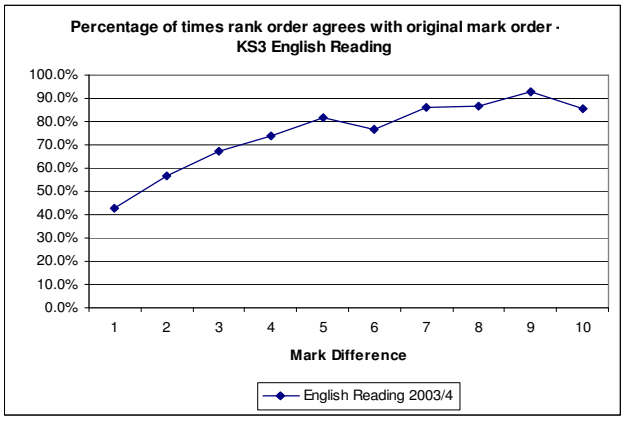
These judgments have similarities with the judgmental part of award meetings but with some important differences. First, as mentioned above, all marks were removed from the scripts we used. In an award meeting the marks are left on. Clearly, the presence of marks on the scripts will have an influence on the judgments being made. Secondly, we asked examiners to look at many (between 15 and 21) different scripts at each mark, compared to only one or two in the award meeting. Thirdly, in an award meeting the nature of the group exercise means that there is potential for discussion and collaboration in making the judgments and also potential for 'peer pressure' to affect a judge's reporting of their judgments. In this research the judges worked individually and their judgments were independent of the other judges.

[1] In the discussion section we consider whether this judgment is really 'absolute' – the distinction is made here to contrast the three different types of judgment made in this study.

With the marks taken off, distinguishing between two scripts which were close together in marks was not going to be easy. Thus, we included the confidence judgment to allow the judges to convey which judgments they found more difficult, and hence might be less likely to get right. We hoped that the confidence judgments would give some interesting insights into the judges' own perceptions of task difficulty.

1.1 Previous work

This study stems from a previous report from Cambridge Assessment (Gill & Bramley, 2006) which used data from rank-ordering studies to look at how far apart in marks scripts had to be before examiners could tell them apart. It was possible to construct a graph of the percentage of script pairs within the rankings that had been 'correctly' ordered[2], for different mark differences. Figure 1 is an example of such a graph, taken directly from the report.



Mark Difference	No of comparisons
1	54
2	159
3	167
4	131
5	77
6	132
7	57
8	45
9	71
10	14

Figure 1: Correct judgments by mark difference (from Gill & Bramley, 2006).

The authors found that while in general the percentage correctly ordered did increase with mark differences, the percentages correctly ordered at the mark differences implied by the zones were mainly between 60 and 65%. However, as these rank-ordering studies were designed to investigate a different question, their design was not ideal from the point of view of the current study. First, there was a lack of true independence in the paired comparisons since they were generated from rankings of ten scripts, and some scripts were seen by the same examiner more than once. Secondly, for some mark differences the number of comparisons was very small.

Therefore, the current research sought to achieve a design which would allow the research question to be answered directly by making all of the judgments genuinely independent paired comparisons as described below in Section 2, ensuring that none of the scripts was seen more than once by any examiner, and by having a sufficient number of comparisons at each mark difference.

Other evidence from research into examiners' judgment of scripts with marks removed suggests this is not a task that they find easy. Forster (2005) asked examiners to make paired comparisons between scripts around key boundaries for GCSE English, A-level Business Studies and A-level Geography. The paired comparison data were then analysed using a Rasch model to come up with a rank order of the scripts in terms of their overall quality. The results showed that the ordering implied by the paired comparisons was generally quite different from the mark rank order. Indeed on only one occasion out of six was there a correlation between the two that was significantly different from zero. He also looked at whether the percentage of correct judgments was higher at larger mark differences. There was some evidence of this,

[2] Whenever we refer to 'correctly judged' or 'correct judgments' in this article we mean that the judged grade was the same as the grade the script actually received (for absolute judgments) or that the order the judges ranked each pair of scripts agreed with the original mark order (for relative judgments).

although the improvements were really quite small and the percentages only just better than chance, even at the maximum difference (four marks).

Similarly, Baird & Dhillon (2005) got examiners to look at scripts around key grade boundaries for GCSE English and A-level Physics. They asked the judges to rank order the scripts they were given (14 scripts, two on each of seven consecutive marks around each boundary) and correlated this order with the original mark order. Again there were few significant correlations. They also looked at to what degree, given the rank order, there was agreement with the original grade classification. There was little agreement, with only one grade boundary out of four giving results much better than chance. They concluded that:

"grade-worthiness judgments, generally speaking, are not fine-grained: they are a broad beamed searchlight for identifying standards … we are utilising examiners' expertise ineffectively by asking them to make fine distinctions between candidates' performances in terms of grade-worthiness". (Baird & Dhillon, 2005, p2).

Several of the studies above attempted to take account of the problem of reliability of marking[3]. Baird & Dhillon (*op. cit.*) only used scripts marked by the principal examiner or where the assistant examiner's marks agreed with the senior examiner. For one of their examples, Gill & Bramley (*op. cit.*) reanalysed the comparison data, but using marks that were generated from a re-marking exercise by just one principal examiner. If the reliability of the original marking was having an effect on the comparisons then the re-marking should have led to better results in terms of percentages correctly ordered. However, despite there being differences between the new and old marks given, there was very little difference in the percentages correctly ordered.

Examples of the use of rating scales for self-reported confidence assessment can be found in Baird (2000) and Baird and Scharaschkin (2002), who both asked judges to rate on a five point scale how difficult they found it to make their judgments of grade-worthiness of scripts (although they did not look at how the ratings given related to how accurate the judgments were). This is slightly different from the confidence judgments used in the present research, as some examiners might perceive that a script was difficult to grade, but in the end feel confident about their judgment. However, it is likely that in general a judgment that is harder to make is likely to be one that the examiner is less confident about.

One premise of this research is that examiners can make relatively quick, 'holistic' judgments about the quality of work in a script (for our purposes a script refers to one exam paper), and can therefore decide which of a pair of scripts is better overall. The research reported here used paired comparisons, but there was no attempt to construct a latent trait of perceived quality out of the judgments (in contrast to much other work on examination comparability that has involved paired comparisons – see Bramley (2007) for a review). The focus was rather on the performance of the judges and the agreement of their judgments with the original marks/grades. It was also specifically designed to avoid one of the weaknesses of the inter-board comparability exercises, where the assumption of independence of the paired comparisons was unlikely to be met because judges saw each script several times and therefore (arguably) might have been influenced by their recollection of them in subsequent comparisons. In the study reported here the judges saw each script *once only*. This aspect of this research marks it out from previous paired comparisons work in this area.

[3] The marking of exam papers is rarely 100% reliable: different opinions on the quality of work and different interpretations of mark schemes lead to different examiners giving quite varied marks to the same script. Thus it is entirely possible that the examiners in studies such as these would disagree with the original mark given on some of the scripts.

2. Method

2.1 Choice of subjects

Two components were chosen for this research, one with a relatively narrow zone and one with a relatively wide zone. However, this was not width in terms of raw marks, but in terms of marks in comparison to the Grade A to Grade E band width. If we assume that the difference in quality of work between an A boundary script and an E boundary script is the same on all papers then each individual mark is worth less (in terms of quality) on a paper with a relatively large A to E band width than it is on a paper with a relatively small A to E band width. Thus, if the zone in terms of *raw* marks is the same on these two papers, the 'true zone' is narrower on the one with the larger A to E band width. The implication of this is that the examiners on the component with the narrower 'true zone' have better 'resolving power' (the ability to detect fine differences in quality) when comparing the quality of work in the sample scripts. One aim of this research was to test the hypothesis that in a subject with a narrower 'true zone' examiners *are* more able to distinguish between scripts a few marks apart (when the marks are removed).

It was thought that the maximum raw mark on a paper might have an impact on the A-E mark range, and so it was decided that the two papers we looked at should have the same maximum mark, but contrasting 'true zone' width. The unit chosen with a narrow zone was an OCR History unit, 2589. It is interesting to note that this is a subject that might be considered to be one of the more difficult to mark consistently – there is more room for subjective judgment in marking this essay-based paper than others such as Maths and Physics, which tend to have more short-answer questions. However, perhaps this is not altogether surprising as it may be that an essay based paper lends itself more to an overall holistic judgment of quality (as happens when setting the zone) than short answer questions (where the judgment of overall quality is likely to involve a summing of individual parts). The unit chosen with a wide zone was OCR Physics 2824.

2.2 Design

The design was necessarily complex as we wanted to have the same number of comparisons at each mark difference, and also to ensure that no one script was seen by any judge more than once. The second of these conditions meant that all of the paired comparisons were truly independent.

We decided to concentrate on comparisons around a grade boundary as this is what happens at an awarding meeting. We used scripts mainly around the A/B boundary, but also included a few C grade scripts as we wanted to have some quite large mark differences in the comparisons. For the History papers we had scripts ranging from 57 marks up to 75 (the A/B grade boundary was at 68 marks and the B/C boundary at 61 marks). For Physics the scripts were from 49 to 67, with the A/B boundary at 60 and the B/C boundary at 53. Apart from the actual marks, the design was the same in both subjects, so that the same number of scripts at each grade and at each mark difference was judged.

Five judges were chosen for each subject. For both, the judges were principal examiners who had been involved in the award meeting for the unit in June 2006. Two of the judges had not previously marked the unit, but as skilled and experienced examiners it was thought they would not find this a difficult task. It was decided that they should undertake the task over a two day meeting, as is the standard for the inter-board comparability exercises. Undertaking repeated paired comparisons is a very monotonous task, and two days is probably the limit of boredom thresholds. This gave a limit on the number of comparisons that would be possible. The number of judgments by each judge at each mark difference is shown in Table 1.

Table 1: Number of judgments at each mark difference.

Mark difference	Judgments by each judge
2	15
4	15
6	15
8	15
10	15
12	5
14	5
16	5

This gave a total of 75 comparisons at mark differences 2-10 and 25 at mark differences 12-16.

Once the set of comparisons for each judge was decided on, each judge's set was sorted into a different random order. This was to ensure that there was no pattern that might be detected in relation to the grades or mark differences.

The scripts were cleaned electronically, by editing scanned images. For the History papers all marks were removed as was all mention of generic band, which would indicate a group of marks. For Physics, all of the numerical marks and ticks were taken off.

At the start of the two-day meeting the examiners were given instructions on the task. We explained that we wanted them to make three judgments for each paired comparison, and to mark these on the recording sheet. For the absolute and relative judgments they needed to make a fairly quick holistic judgment about the quality of work in each script, taking into account question difficulty, much as they would in an award meeting. Tied judgments were not allowed. The judges were not given any clues about the range of marks or grades that the scripts were given as it was felt that this might influence them – for instance there might be a script that an examiner thought worthy of a grade D, but if we had already told them that the range of grades was only A to C then clearly they would not give it a D. We made it very clear that they should not re-mark the scripts. We also asked them to give a confidence rating about the absolute and relative judgments, on the following scale: 1 = Not confident, 2 = Fairly confident, 3 = Very confident. An example of the recording sheet is in Appendix 1. Finally, the timings were explained to them so they knew how fast to work.

3. Results

We initially recruited five judges for each subject, and anticipated that they would be able to complete 90 paired comparisons each, giving a total of 450 paired comparisons in each subject. However, at the History meeting it became clear that due to the nature of the paper chosen the examiners were not going to complete all of their judgments. Another examiner was recruited to do at home many of the judgments for which the other examiners did not have time. In all 413 comparisons were made.

We begin the results section with some analysis of the raw data for each of the three types of judgment. Then we go on to consider a possible statistical model for assessing which factors were most important in determining the accuracy of the judgments.

3.1 Absolute judgments

Table 2: Absolute judgments by grade.

	History			Physics		
Grade	Incorrect	Correct	% correct	Incorrect	Correct	% correct
A	197	131	39.9	278	76	21.5
B	212	132	38.4	292	84	22.3
C	97	57	37.0	106	64	37.6
Total	506	320	38.7	676	224	24.9

For History, the overall level of agreement with the original grade was below 40%. There was little evidence that any grade was easier to judge, with A very slightly better than B and B very slightly better than C. The overall level of agreement in Physics was substantially worse than with History: only 25% of the judgments agreed with the original grade the scripts received. The percentage correct was considerably higher with C grade scripts than with either grade A or B.

Table 3: Absolute judgments by judge.

	History			Physics		
Judge	Incorrect	Correct	% correct	Incorrect	Correct	% correct
1	127	41	24.4	134	46	25.6
2	108	72	40.0	100	80	44.4
3	74	28	27.5	144	36	20.0
4	68	82	54.7	129	51	28.3
5	52	42	44.7	169	11	6.1
6	77	55	41.7			

In both subjects there were some substantial differences between judges. In History, judge 4 performed the best, whilst 1 and 3 performed a lot worse than the others. In Physics, judge 2 performed substantially better than the other judges, whilst judge 5 performed much worse. It is worth noting that in both subjects the judge who performed the best was the Principal Examiner (PE) for that unit. As such, they might be expected to have more knowledge and experience of the unit and of which features are important when judging grade-worthiness.

Table 3 has no information on how far the incorrect grade judgments were from the true grades. Appendix 2 shows the distribution of judged grades within each true grade for each judge. It is clear from these graphs that there was quite a spread of judged grades within each true grade and often the judges gave grades that were lower than the true grades. Only on a few occasions was the judged grade higher than true grade. This seems to be more consistently the case in Physics than History. Judges 4 and 5 in History at least confined themselves almost exclusively to A-C grades. In Physics, only judge 2 seemed to have any substantial agreement with the true grade.

Thus in both subjects, but particularly Physics, there seems to be a discrepancy between the standard as set at the award meeting and the judgment of the quality of work by the individual judges. As several different factors feed into determining the grade boundary at an award meeting we have no way of knowing the relative influence of each factor.

It may be that the differences were partly due to not informing the examiners of the range of grades they were looking at: if they guessed that there were at least some Ds and Es in their allocation then the lowest quality scripts would have to be these, even if they really judged them as Cs or above. This would have the effect of dragging the means down somewhat. This shows how difficult it is in a study such as this to control all relevant factors – in this instance the prior beliefs of the judges about what they were judging.

3.2 Relative judgments

In History the examiners correctly ordered 66.1% of the paired comparisons. In Physics the figure was 77.8%. Figure 2 shows the percentage correct for each mark difference:

History *Physics*

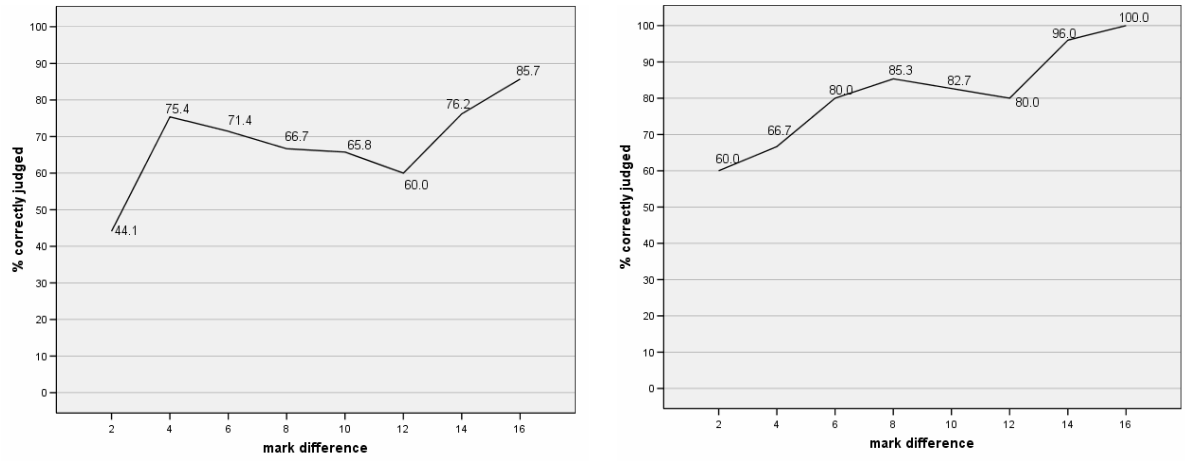


Figure 2: Percentage of pairs correctly ordered, by mark difference.

We would expect the percentage correctly ordered to increase as the mark difference increased, but in History this was not consistently the case. The examiners clearly performed a lot better with a difference of four marks than with a two mark difference. However, at 6, 8,10 and 12 mark differences the percentage correct fell compared to the previous mark difference! It was only at a 14 mark difference that the percentage increased. See the discussion section for further comment on this surprising result. There was a far more consistent pattern with the Physics paper with mainly increasing success at larger mark differences. The only exceptions were at 10 and 12 marks different where the percentage fell compared to 8 marks different. However, these were all above 80% correct so were already very high.

Table 4: Relative judgments by judge.

	History			Physics		
Judge	Incorrect	Correct	% correct	Incorrect	Correct	% correct
1	26	58	69.0	29	61	67.8
2	37	53	58.9	10	80	88.9
3	18	33	64.7	31	59	65.6
4	21	54	72.0	18	72	80.0
5	16	31	66.0	12	78	86.7
6	22	44	66.7			
Total	140	273	66.1	100	350	77.8

For History the differences among the judges were not large, varying between 59% and 72%. In Physics there was slightly more variation: judges 1 and 3 performed substantially worse than the other three judges. Once again the examiner who performed the best in each subject was the PE for the unit. Overall the percentage of correct relative judgments was much higher than for absolute judgments, but now the Physics judges had a slightly higher success rate than the History judges.

3.3 Confidence judgments

Table 5 shows the mean confidence rating scores of each judge for each of the types of judgment (1 = Not confident, 2 = Fairly confident, 3 = Very confident.)

Table 5: Mean confidence rating score on each type of judgment.

	History			Physics		
	Absolute	Relative	Overall	Absolute	Relative	Overall
Judge 1	2.18	2.43	2.26	1.93	2.37	2.08
Judge 2	2.21	2.48	2.30	1.79	2.56	2.04
Judge 3	1.97	2.18	2.04	1.53	1.64	1.57
Judge 4	2.46	2.57	2.50	1.26	1.98	1.51
Judge 5	2.70	2.39	2.59	1.71	2.11	1.84
Judge 6	1.90	2.09	1.96			
Overall	2.22	2.38		1.65	1.65	

A few points are of interest here. First, it seems that overall the History judges were more confident about their judgments than the Physics judges. The overall means for both ratings were higher for History, and most History judges had an overall mean higher than the Physics judges. However, we should be cautious about concluding too much from this because different judges may have perceived confidence (and the confidence rating scale) in a different way.

What we can more safely conclude is that whilst the History judges rated their level of confidence to be roughly the same for both judgments, in Physics the examiners were far more confident about the relative judgment of which script was better than the absolute judgments of the grades. Inspection of the previous Tables 3 and 4 shows that this extra confidence was justified.

Next we looked at the percentage of correct judgments for each rating. We might expect a higher percentage correct when the judges were more confident:

Table 6: Percentage of correct judgments for each rating score (relative judgments).

	History			Physics		
Rating	1	2	3	1	2	3
Judge 1	75.0*	57.5	80.0	68.4	55.6	73.1
Judge 2	50.0*	40.7	69.8	83.3	68.8	95.2
Judge 3	41.7*	72.2	71.4	58.5	66.7	88.9
Judge 4	25.0*	75.0	74.5	70.4	76.3	96.0
Judge 5	50.0*	50.0*	76.9	87.5	83.3	92.3
Judge 6	50.0*	66.7	77.8			
All	48.0	60.5	74.6	69.6	73.0	87.9

* = Low n, less than 20.

In History it is difficult to read anything into the 'not confident' rating, due to the very low number of occasions that the judges chose that option. Most of the judges did perform better when they

felt 'very confident' compared with feeling 'fairly confident', although the performance of judges 3 and 4 was essentially the same for both ratings.

For each of the Physics judges the percentage correct when they were 'very confident' was clearly higher than 'fairly confident' and 'not confident'. However, the judges did not really perform any better when they were 'fairly confident', compared with 'not confident'.

Table 7 gives the percentages correct for the absolute judgments of grade, for each of the confidence ratings.

Table 7: Percentage of correct judgments for each rating score (absolute judgments).

	History			Physics		
Rating	1	2	3	1	2	3
Judge 1	23.1*	20.5	34.9	19.4	20.3	73.7
Judge 2	15.0	37.9	52.6	41.5	35.2	81.5
Judge 3	13.3*	23.0	66.7*	21.4	18.9	50.0*
Judge 4	33.3*	50.7	60.6	27.8	28.6	100.0*
Judge 5	n/a	39.3	46.9	10.5	4.2	0.0*
Judge 6	31.4	40.0	63.6			
All	23.3	33.8	52.0	25.3	19.5	70.4

* = Low n, less than 20.

The trend in History is fairly clear, with judges performing better when 'fairly confident' than 'not confident' (although note the low numbers of 'not confident' judgments) and better still when 'very confident'. However, the overall percentage correct for judgments where the judge was 'very confident' was only just over 50% – not a high figure considering their apparent level of confidence.

For Physics, only two of the judges had a reasonable number of judgments where they felt 'very confident' and for these they performed much better than the 'fairly' or 'not confident' judgments. However, there was no relationship between the percentages correct when the judgments were 'not confident' or 'fairly confident'.

Overall, there is some evidence here of a relationship between the judges' confidence and their actual performance. However, this did seem to be limited to occasions when the judges felt 'very confident' about a judgment. There was little evidence that they performed better when they felt 'fairly confident' as opposed to 'not confident'.

3.4 Logistic regression

In the analysis above we considered several factors that may have had an impact on the absolute and relative judgments being made, some of which had a clear effect and some which did not. We can test statistically for each of the factors, and explore more complex interactions which may not have been apparent, using logistic regression. Logistic regression is useful when the variable being predicted is binary – i.e. can only take two values. In our case we are interested in whether the judgment of the better script or of the grade of the script was 'correct' or not. It is then possible to look at different combinations of the independent variables and observe how the probability of a correct judgment changes.

Logistic regression uses a log-odds transformation of the dependent variable. This is to avoid the possibility of predicting probability values less than zero or greater than one. Once the value of the dependent variable has been predicted this can be transformed back into a probability. The general form of all the models described here is:

$$\ln\left(\frac{P}{(1-P)}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 \qquad \text{etc.}$$

where P is the probability of a 'correct' judgment, α is the intercept, $\beta_1 \ldots \beta_n$ are regression parameters, and $x_1 \ldots x_n$ are the independent variables which in this case were mark difference, judge, confidence rating etc[4].

3.4.1 Absolute judgments

The independent variables included in the modelling were correct grade (A,B,C – coded 6,5,4), judge (1 to 5/6) and confidence rating (1-3).

In History, the significant predictors of success in making a correct judgment of grade were judge and confidence rating. The correct grade of the script was not relevant (i.e. the judges were no more likely to make a correct judgment about grade A scripts than they were about B or C scripts.). There were no significant interactions.

Table 8: History absolute judgments – summary of significant effects.

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
judge	1 vs 6	0.373	0.223	0.623
judge	2 vs 6	0.767	0.477	1.232
judge	3 vs 6	0.486	0.274	0.862
judge	4 vs 6	1.224	0.741	2.020
judge	5 vs 6	0.677	0.378	1.212
confid_abs1	1 vs 3	0.280	0.154	0.510
confid_abs1	2 vs 3	0.512	0.368	0.712

The magnitude of the effect of individual variables can best be assessed via the point estimate of the odds ratio. This is the change in the odds of a correct judgment for a unit increase in the variable in question. For categorical variables we require a base category and the odds ratio then tells us how the odds of a correct judgment change compared with the base category. The base categories in Table 8 were Judge 6 and confidence rating 3.

The estimate for confidence rating of 1 was significantly different from 3, with a value of 0.280 meaning the odds of correct judgment when the rating was 1 were around a quarter of the odds with a rating of 3. The estimate for a confidence rating of 2 was also significantly different from 3: its value was 0.512, meaning that the odds of correct judgment with a rating of 2 fell by around a half compared with a rating of 3.

Similarly, there were clear differences between the judges, with the odds of a correct judgment for judges 1 and 3 being less than half those for judge 6, and no overlap in the confidence limits for judges 1 and 4. These findings support the earlier interpretations of the raw data in Tables 2, 3 and 7.

In Physics, in contrast, the correct grade of the script was also significant, in addition to the judge and confidence. Again there were no significant interactions.

[4] Effects of individual scripts were treated as part of the random error, since no script could be seen more than five times

Table 9: Physics absolute judgments – summary of significant effects.

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
script1_grade 4 vs 6	2.609	1.662	4.097
script1_grade 5 vs 6	1.266	0.858	1.866
judge 1 vs 5	4.886	2.371	10.067
judge 2 vs 5	11.298	5.590	22.838
judge 3 vs 5	4.377	2.104	9.106
judge 4 vs 5	6.867	3.304	14.273
confid_abs1 1 vs 3	0.132	0.064	0.272
confid_abs1 2 vs 3	0.117	0.059	0.232

The odds of correct judgment were significantly greater for grade C compared to grade A (2.61 times). For the confidence rating both variables were highly significant: the odds of success for a rating of 1 were about 1/8th the odds with a rating of 3 and for a rating of 2 about 1/10th the odds with a rating of 3. The odds ratio estimates appear very large for the judges, but this was because the baseline judge (judge 5) had a very low odds of making a correct judgment. It is clear that all the judges were better than judge 5, and that judge 2 was significantly better than judges 1, 3 and 5. Again, this more complex analysis supports the earlier interpretation of Tables 2, 3 and 7.

3.4.2 Relative judgments

The independent variables included in the modelling were the mark difference, the grade combination of the two scripts, whether they answered the same question (History only), and the confidence rating.

In History, there was no evidence that the grade combination or the questions answered had any effect on the probability of a correct judgment. There were no significant interactions.

Table 10: History relative judgments – summary of significant effects.

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
MARKDIFF	1.082	1.024	1.144
confid_rel 1 vs 3	0.295	0.154	0.563
confid_rel 2 vs 3	0.504	0.320	0.794

The mark difference coefficient was significant: each increase of one mark improved the odds of a correct relative judgment by a factor of 1.082. However, the model including mark difference did not fit the data well[5], which is not surprising given the shape of the plot in Figure 2. It is clear that the probability of a correct relative judgment did not increase monotonically with increasing mark difference.

For the confidence ratings, pairs of scripts judged with a rating of 1 had odds of success reduced by about three and a half times compared with rating 3. The odds for scripts rated 2 were around half that of scripts rated 3.

[5] As diagnosed by the Hosmer & Lemeshow (2000) goodness of fit test – see SAS online documentation http://support.sas.com/documentation/cdl/en/statug/59654/HTML/default/statug_logistic_sect039.htm .

Interestingly, the differences between the judges were no longer significant. This point is explored further in the discussion.

The results in Physics were more complex. There was no significant effect due to the grade combination. The mark difference coefficient was significant, and this time the model fitted well – Figure 2 shows a good increasing relationship between mark difference and probability of correct judgment. However, there was a significant interaction between mark difference and confidence – for confidence ratings of 2 and 3 the mark difference had a much stronger relationship to the probability of success than it did for a confidence rating of 1. This is shown in Figure 3 , which shows the modelled probabilities of success (not the raw data, the graph of which would obviously be less smooth).

Furthermore, unlike History, the judges did differ in their chances of making a correct relative judgment.

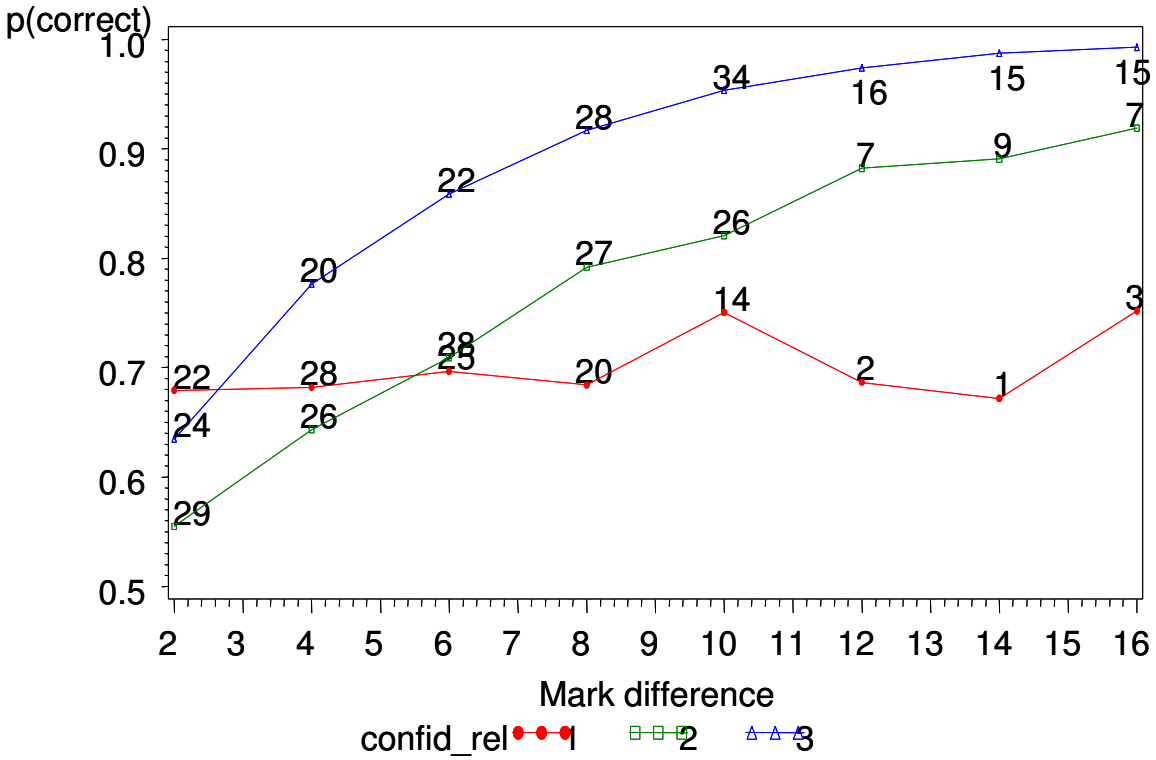


Figure 3: Physics relative judgments – modelled probability of correct judgment (data points labelled with the number of cases contributing to that point).

Table 11: Physics relative judgments – summary of significant effects.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.1929	0.2559	0.5683	0.4509
MARKDIFF		1	0.1809	0.0396	20.8799	<.0001
confid_rel	1	1	0.5849	0.3678	2.5288	0.1118
confid_rel	2	1	-0.4564	0.3457	1.7431	0.1867
MARKDIFF*confid_rel	1	1	-0.1363	0.0548	6.1875	0.0129
MARKDIFF*confid_rel	2	1	0.0101	0.0505	0.0398	0.8418
JUDGE	1	1	-0.8134	0.2446	11.0557	0.0009
JUDGE	2	1	0.5910	0.3128	3.5694	0.0589
JUDGE	3	1	-0.5734	0.2347	5.9680	0.0146
JUDGE	4	1	0.1030	0.2543	0.1642	0.6853

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
JUDGE	1 vs 5	0.222	0.095	0.516
JUDGE	2 vs 5	0.903	0.342	2.384
JUDGE	3 vs 5	0.282	0.126	0.630
JUDGE	4 vs 5	0.554	0.240	1.284

Note that it is not possible to estimate the odds ratio for the effect of the mark difference and the confidence ratings because of the significant interaction between the two. Table 11 therefore includes the model parameter estimates for completeness. Table 11 shows that Judge 5's relative judgments were more likely to be correct than those of any other judge, especially judges 1 and 3.

In general, the results of the logistic regression modelling back up the interpretations of the raw data in Section 3.3, adding some extra insights about the significance of different effects and the interaction (in Physics relative judgments) between mark difference and confidence.

4. Discussion

The process of awarding should involve examiners' professional judgment on the quality of the work seen in the sample scripts. As the QCA Code of Practice states, "Each boundary must be set using professional judgment. The judgment must reflect the quality of the candidates' work, informed by the relevant technical and statistical evidence" (QCA 2006, p35).

However, it can be argued that the professional judgments made are not 'pure' judgments of quality – they have been affected by the presence of the marks on the script and other factors. For instance, the range of scripts they look at is determined prior to the award meeting by the PE, and must take account of the relevant technical and statistical information. It is possible for the examiners to ask to see scripts outside of this range, if they genuinely feel the standard lies elsewhere, but would they feel able to do this? It might cause conflict with the other examiners and the Chair, and would certainly prolong the meeting. Previous research has found that some examiners might feel under at least some pressure not to 'rock the boat' too much (see Murphy *et al* 1996; Cresswell, 1997).

By the same token it is also plausible that the examiners will feel more inclined to choose a zone in the middle of the range of scripts they look at, than at the edge. In short, it is arguable that the whole nature of the mandated awarding procedure directs the examiners towards a particular mark range.

We note that there were several important differences between the judging situation in this research and the situation in an award meeting. First, the examiners in this research were unaware of the mark given to each script, and of where the scripts were in comparison to the range, and therefore where the boundary was likely to lie.

Second, there were no archive scripts available to remind examiners of the standard. Baird (2000) investigated the accuracy of grading judgments depending on whether archive scripts were used in English Literature and Psychology A-level papers. The major difference from our research was that she left the marks on the papers. The use of archive scripts had no impact on the accuracy in English Literature, but in Psychology the group of examiners who had 'balanced' E grade archive scripts performed significantly better than the examiners who had no archive scripts, unbalanced E grade archive scripts or D grade archive scripts. It may be that using archive scripts in our study would have made a difference to the accuracy of the absolute judgments.

Third, the awarding process relies on the *collective* professional judgment of the examiners. We asked the examiners to grade each script individually, and they did not discuss their decisions with anyone else (although there was occasional clarification of mark schemes). In an award meeting the decision on the zone is by consensus and this may lead to a group dynamic effect, where if all other judges have called it 'in', it is a lot more difficult to go against all of them even if that is what the judge believes. Award meetings can differ in the way the judges are asked to convey their decision of grade-worthiness, so this is not to say this effect was present at the meetings for the units used in this research. This is an area of further research that could be explored.

Finally, as mentioned earlier in this report, the decision not to inform the examiners of the range of grades of the scripts they looked at may have had an effect. In an award meeting, they know that the script they are looking at will be one of two grades. Here, they had free rein to decide on any grade. If they assumed that the scripts used in the study would cover the full range (A-U) they would have felt the necessity to include some lower grades in their judgments. However, some of the examiners at least did not feel this. One History examiner judged all scripts to be in the range A to C and another only awarded a lower grade on one occasion. Moreover, had we told the examiners the true range, we would arguably have been directing their judgments to a significant extent.

With these differences duly noted, this research has demonstrated that examiners often cannot grade the scripts in a manner consistent with the decisions made at the award meeting. Remember, the judges in this exercise were the same ones that came up with the boundaries in the award meeting. In History, less than 40% of the absolute judgments of grade were correct, and the figure in Physics was less than 25%. Furthermore, there was no sign that the judges were better at judging grade A scripts, which is what we might expect if they have, as the procedures apparently assume, an internal standard representing a grade A. One has to question, therefore, the usefulness of the examiners undertaking a task they seem unable to do accurately when the marks are removed. We are told that professional judgment is integral to the process of grading, but it seems likely that other factors have a considerable (and perhaps overriding) influence on these judgments.

The performance of the examiners in making the relative judgments was much better. 66.1% of the History paired comparisons and 77.8% of the Physics comparisons were correctly ordered. This finding provides further support for the claim that the best way to use expert judgment in this kind of decision making is via relative judgments and not absolute judgments – a case which has most often been made in the area of comparability research (e.g. Pollitt & Elliott, 2003; Pollitt, 2004; Bramley, 2007), but also in the area of awarding (e.g. Baird & Dhillon, 2005; Black & Bramley, 2008).

Why is it that the judges were so much better at making relative, rather than absolute, judgments in this context? The simple answer might be that in the relative judgment they had only two options, script 1 or 2, whereas in the absolute judgment they had six grades to choose between.

Thus if they were guessing they were more likely to guess correctly for the relative judgments. There may also be another reason: calling the grade judgments 'absolute' judgments is a bit misleading because, as Laming puts it: "There is no absolute judgment. All judgments are comparisons of one thing with another." (Laming 2004, p9). The examiners were making their judgments about the grade by comparing the quality of the script to some sort of internal standard of what an A grade script should look like. As expert examiners they are supposed to internalise the standard before an award meeting (and to reinforce it by inspecting the archive), to enable the process of 'absolute' judgments of grade. It seems likely therefore that it is easier to make the relative judgment of comparing two physical scripts in front of you than make the relative judgment between one script and this internal standard.

It is worth considering the percentage of correct judgments at the mark difference implied by the zone. The zone for History was 67-69, thus a mark difference of three should have been enough to distinguish between scripts. For Physics the zone was 63-58, so a difference of six marks was enough. In our exercise the History examiners managed 44.1% accuracy on relative judgments of pairs of scripts separated by two marks, and 75.4% at four marks. In Physics they achieved 80% at six marks different. However, the History percentage decreased successively at six, eight, ten and twelve marks different, ending up at only 60%. This suggests there is some doubt that the examiners could consistently achieve the 75% figure at a mark difference of four.

We proposed earlier that in the History unit, with a narrow zone, it should have been easier to distinguish between scripts a few marks apart than in Physics, with a wider zone. Figure 4 compares the two units:

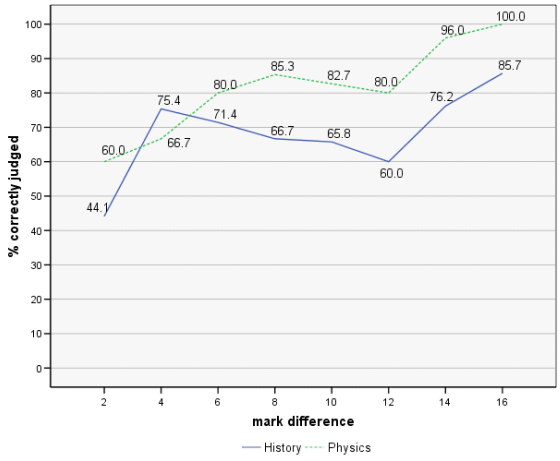


Figure 4: Relative judgments – percentage correct by mark difference for both units.

We can see that the hypothesis has not been supported. For all mark differences apart from 4 marks, the judges on the Physics paper achieved a higher percentage of correct relative judgments. It is likely therefore that there are other factors that had an impact on the judges' ability to choose between the scripts. It may be that the cognitive load involved in making a holistic judgment was greater in History. Several of the judges commented that the paper chosen was difficult to judge because it involved three different types of questions; although, since more of the *absolute* judgments were correct in History this is perhaps a false supposition. Perhaps more likely is the effect of the reliability of the marking of the scripts. Previous research has shown (see Murphy 1978, 1982; Newton, 1996) that the marking of papers with long essay type questions (such as History) is less reliable than papers with short answer, objective questions (such as Physics). Thus, in History the order of each pair of scripts as defined by the original marks is more likely to have been "incorrect", particularly at low mark differences. Indeed, the difficulty of agreeing on a single mark in a more subjective subject like History shows that the concept of a 'correct' mark for a script is somewhat fuzzy. However, it is arguable that the cognitive processes involved in marking a History script are more similar to those involved in

making a judgment about grade-worthiness than they are in Physics. This could explain why the absolute judgments in History were more accurate than those in Physics.

In terms of the confidence ratings several of the History judges felt the task involved in judging this particular unit was difficult, yet they still rated themselves as more confident in both tasks on average than the Physics judges. They were also fairly equivalent in their confidence about both types of judgment, whereas the Physics judges were clearly more confident about the relative judgments than the absolute. The difference between the average confidence rating when they made the correct judgment and when they got the judgment wrong was generally small, suggesting that the judges did not have much awareness of their own judgmental skills. However, the interaction between confidence and mark difference in the Physics relative judgments is interesting – Figure 3 shows that the effect of increasing mark difference had little impact on the modelled probability of a correct relative judgment when the judges were not confident. It would be interesting to see if there were any particular features in common of the scripts involved in those 'not confident' relative judgments. If so, one implication might be not to involve that kind of script in an award meeting.

Finally we note some limitations with this research. The choice of the History unit caused some difficulties with the judgmental task, due to the complexity of judging the various merits of three different types of questions with different mark tariffs, and the possibility of question choice in the rubric. We could counter this by saying that the judgments made at award meetings need to be done for this unit in the same way as any other, but perhaps future research could concentrate on papers that would be easier to judge, particularly given the number of judgments that were expected in a relatively short amount of time.

As described above, the conditions of our experiment were different from an award meeting in many respects. The marks were removed from the scripts, there were no archive scripts, there was no Chair to influence the outcome, the judges looked at a larger number of scripts over a larger range of marks and made their judgments without conferring with colleagues.

In conclusion, while acknowledging that there were significant differences between the judgmental situation in our study and that in an award meeting, we would argue that the conditions in this study are preferable from the point of view of obtaining a 'pure' judgmental outcome. The fact that the judges were more accurate at making relative, rather than absolute judgments raises some doubts as to whether the current awarding procedures make best use of expert judgment. Alternatives such as rank ordering, or paired comparisons, which require relative rather than absolute judgments, may be a better way of using the examiners' skills, particularly if this is done independently of the external statistical information about mark distributions and cohort characteristics.

References

Baird, J-A. (2000). Are examination standards all in the head? Experiments with examiners' judgments of standards in A level examinations. *Research in Education,* 64, 91-100.

Baird, J-A. & Dhillon, D. (2005). Qualitative expert judgments on examination standards: valid, but inexact. AQA research report RPA_05_JB_RP_077. AQA: Guildford.

Baird, J-A. & Scharaschkin, A. (2002). Is the Whole Worth More than the Sum of the Parts? Studies of Examiners' Grading of Individual Papers and Candidates' Whole A-level Examination Performances. *Educational Studies,* 28, 143-162.

Black, B. & Bramley, T. (2008). Investigating a judgmental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education, 23(3)*, 357-373.

Bramley, T. (2007). Paired comparison methods. In P.E. Newton, J. Baird, H. Goldstein, H. Patrick and P. Tymms (Eds.). *Techniques for monitoring the comparability of examination standards*. London. Qualifications and Curriculum Authority.

Cresswell, M.J. (1997). *Examining judgements: theory and practice of awarding public examination grades*. PhD Thesis, Institute of Education, University of London.

Forster, M. (2005). Can examiners successfully distinguish between scripts that vary by only a small range on marks? Unpublished internal paper, Oxford Cambridge and RSA Examinations

Gill, T. & Bramley, T. (2006). Can examiners distinguish between scripts a few marks apart? Evidence from rank ordering studies. Cambridge Assessment internal report

Hosmer, D. W., Jr. and Lemeshow, S. (2000). *Applied Logistic Regression*, Second Edition, New York: John Wiley & Sons.

Laming, D., (2004) Human Judgment: the Eye of the Beholder. London, Thomson

Murphy, R.J.L. (1978). Reliability of Marking in Eight GCE Examinations. *British Journal of Educational Psychology.* 48, 196-200.

Murphy, R.J.L. (1982). A Further Report of Investigations into the Reliability of Marking of GCE Examinations. *British Journal of Educational Psychology.* 52, 58-63.

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.

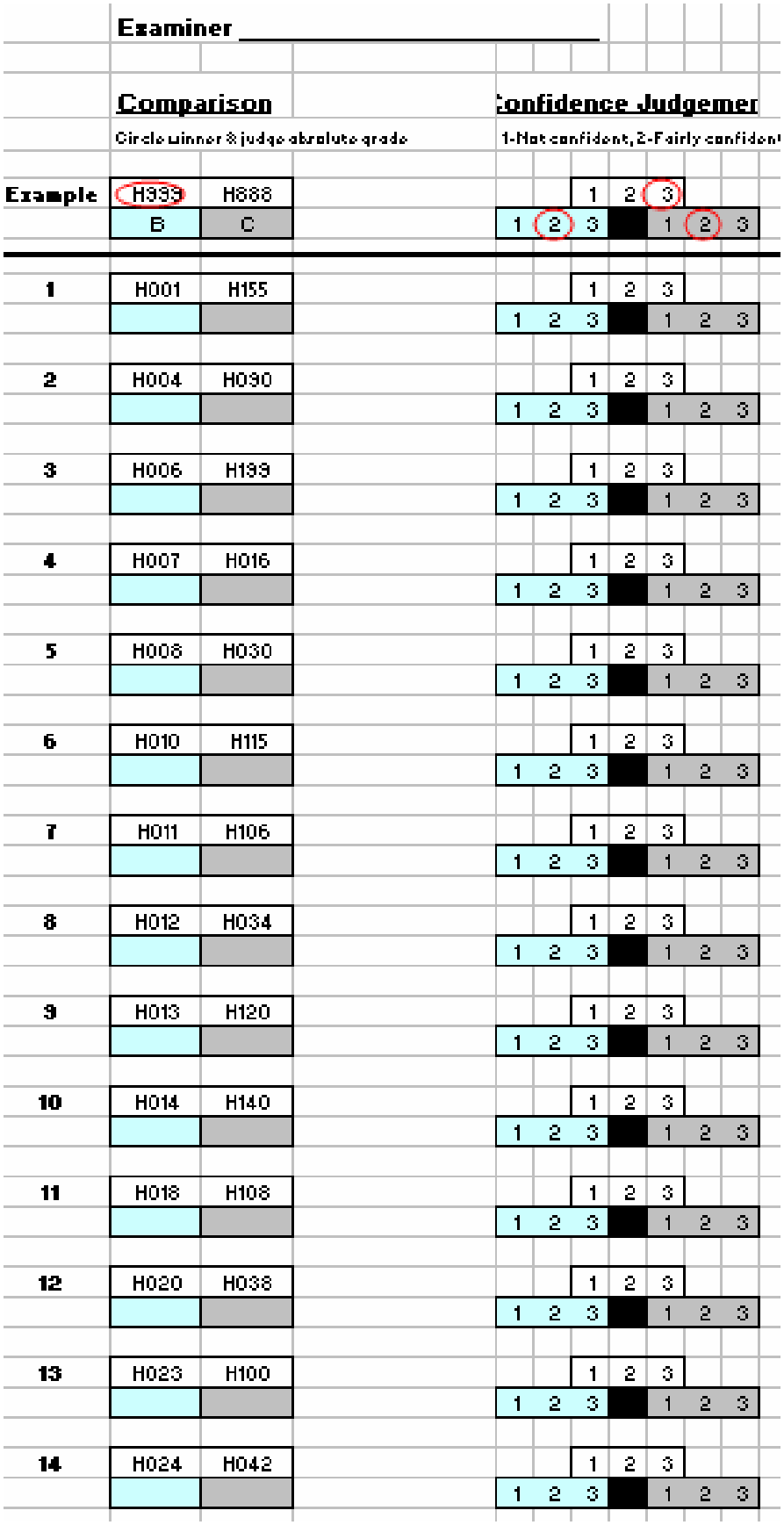
Newton, P.E. (1996). The Reliability of Marking of General Certificate of Secondary Education Scripts: Mathematics and English. *British Educational Research Journal,* 22(4), 405-420.

Pollitt, A & Elliott, G. (2003) *Finding a proper role for human judgment in the examination system.* Paper given at the QCA 'Comparability and Standards' seminar, Newport Pagnell, 4th April. Available at http://www.cambridgeassessment.org.uk/research/confproceedingsetc Accessed 22/06/06.

Pollitt, A. (2004). *Let's stop marking exams.* Paper presented at the annual conference of the International Association for Educational Assessment (IAEA), Philadelphia.
Available at http://www.cambridgeassessment.org.uk/research/confproceedingsetc Accessed 22/02/06.

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

Appendix 1: Example recording sheet



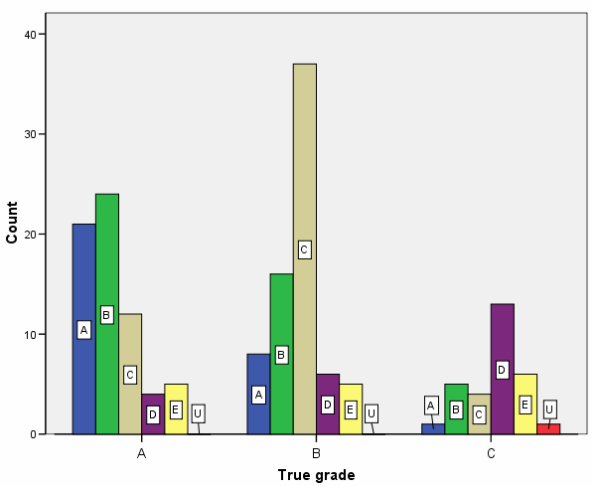
Examiner _____

	Comparison		Confidence Judgemer	
	Circle winner & judge absolute grade		1-Not confident, 2-Fairly confident	
Example	H333	H888	1 2 3	
	B	C	1 2 3	1 2 3
1	H001	H155	1 2 3	
			1 2 3	1 2 3
2	H004	H090	1 2 3	
			1 2 3	1 2 3
3	H006	H139	1 2 3	
			1 2 3	1 2 3
4	H007	H016	1 2 3	
			1 2 3	1 2 3
5	H008	H030	1 2 3	
			1 2 3	1 2 3
6	H010	H115	1 2 3	
			1 2 3	1 2 3
7	H011	H106	1 2 3	
			1 2 3	1 2 3
8	H012	H034	1 2 3	
			1 2 3	1 2 3
9	H013	H120	1 2 3	
			1 2 3	1 2 3
10	H014	H140	1 2 3	
			1 2 3	1 2 3
11	H018	H108	1 2 3	
			1 2 3	1 2 3
12	H020	H038	1 2 3	
			1 2 3	1 2 3
13	H023	H100	1 2 3	
			1 2 3	1 2 3
14	H024	H042	1 2 3	
			1 2 3	1 2 3

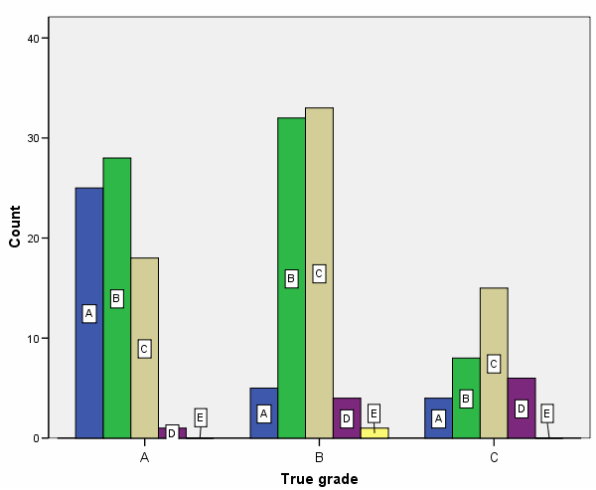
Appendix 2: Distribution of judged grade within each true grade

History

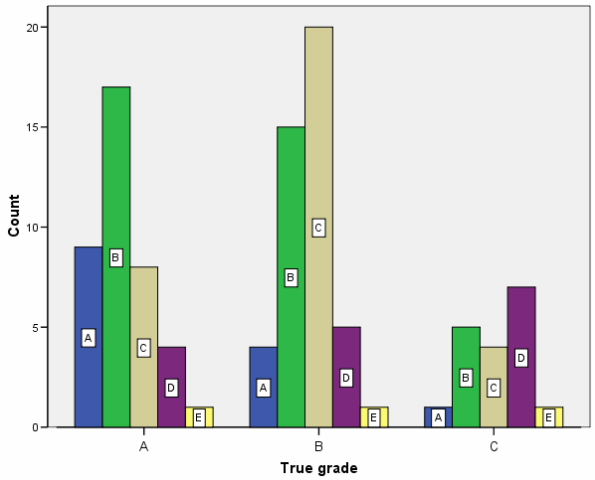
Judge1



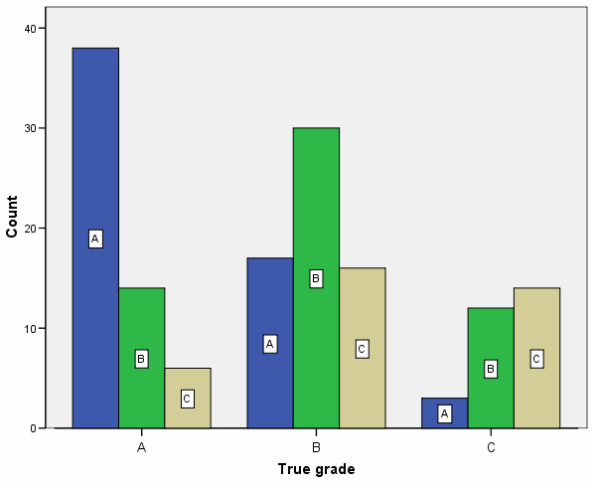
Judge 2



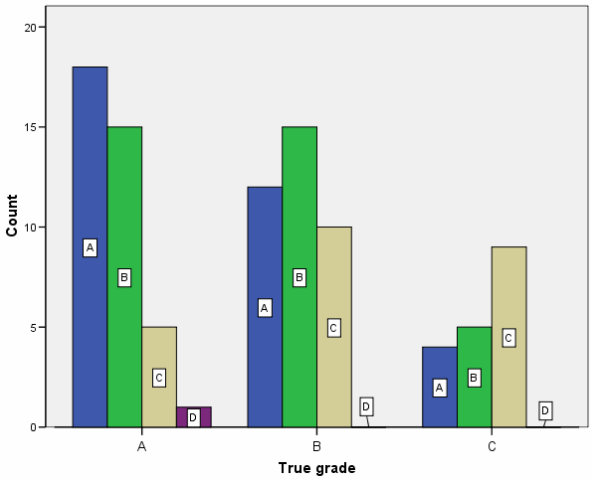
Judge 3



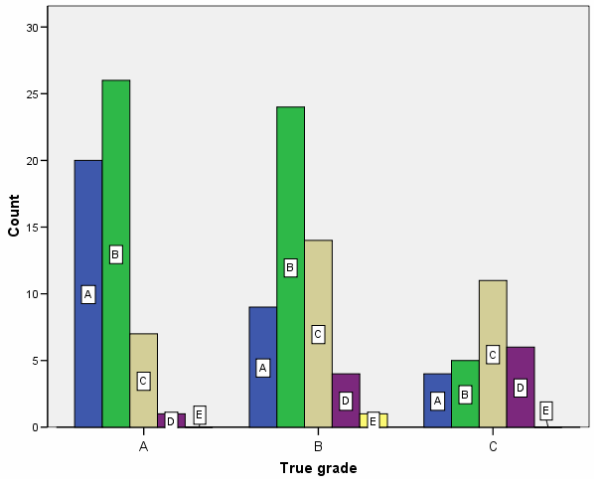
Judge 4



Judge 5

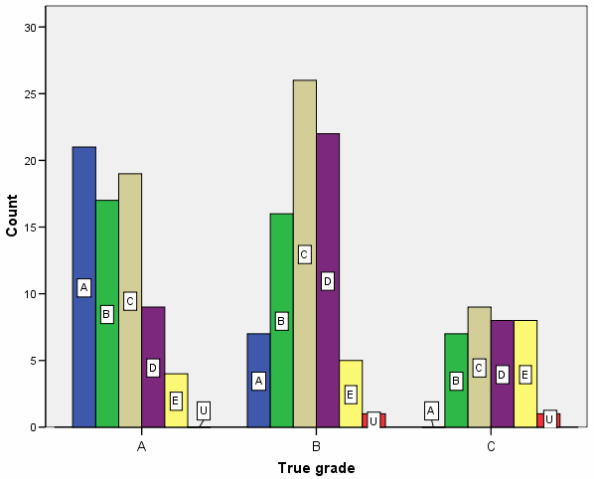


Judge 6

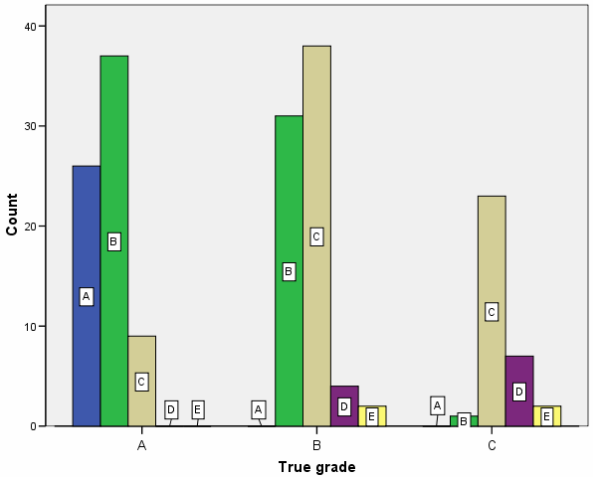


Physics

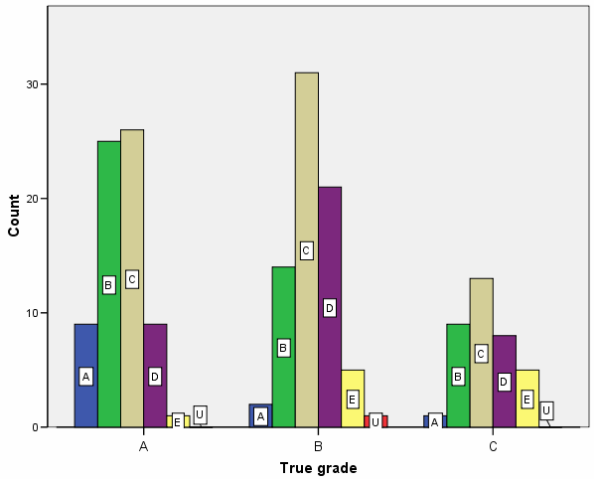
Judge1



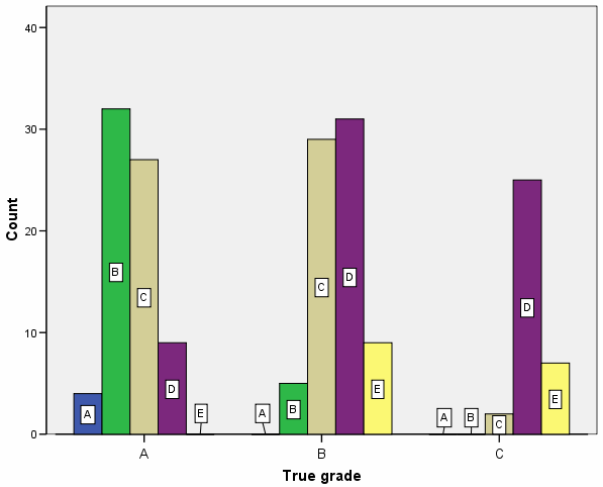
Judge 2



Judge 3



Judge 4



Judge 5