Title: Comparing the reliability of standard maintaining via examiner judgement to statistical approaches

Author: Tom Benton
Principal Research Officer, Cambridge Assessment
Email address: Benton.T@cambridgeassessment.org.uk

Abstract: This paper compares the reliability of two methods for maintaining examination standards - a qualitative comparison of scripts in different years by expert judges, and the use of statistical predictions. Initially, the paper will introduce a theoretical model for the way in which expert judgement operates. This model encompasses the relationship between the score awarded to an examination script and the perceived holistic quality of the script by an expert judge, as well as the expected level of inter-judge variation in the level of perceived holistic quality for a single script. The paper will then demonstrate that this model provides results that are consistent with existing research, including both studies that are critical and studies that are supportive of the use of expert judgement. The model will then be applied to examine how the reliability of expert-driven approaches to standard maintaining is dependent upon the number of judges involved and the number of candidate scripts that are scrutinised. Finally the expected reliabilities of grade boundaries derived via differing approaches to expert judgement will be compared to a purely statistical approach to the same problem.

Keywords: Standard maintaining, comparative judgement, reliability

Introduction

For most examinations in England, once a set of scripts have been marked, and thus assigned numerical scores, it is necessary to determine how the scores should be divided into grades. This process of grade awarding is done with the intention that grades should reflect the same standards over time. In England, grade awarding has traditionally relied upon two distinct sources of information: statistical evidence and expert judgement.

In its most simple form statistical evidence may simply present the percentage of candidates achieving each grade in the previous exam session. This information may be combined with an assumption that the ability of successive cohorts of candidates will not differ too dramatically to guide awarding. More recently, statistical evidence has taken the form of predicted achievement of the cohort of candidates in the current year based upon their prior attainment and upon statistical models developed using historical data.

Expert judgement, on the other hand, attempts to directly assess the quality of scripts. It may be applied in a number of different forms including directly verifying the grade worthiness of scripts with different numbers of marks or comparative judgement where scripts with different numbers of marks may be compared to "benchmark" scripts that were deemed to possess minimally sufficient quality to be awarded a given grade in the past.

Some previous research has suggested that comparative judgement may provide a reliable mechanism to maintain standards (for example, Novakovic and Suto, 2010). However, other research is more critical of such approaches (for example, Stringer 2012). However, very little existing research has explicitly tried to estimate the reliability of methods for standard maintaining purely based upon comparative judgement.

This paper examines the potential reliability of a standard maintaining technique based upon comparative judgement and compares this to the expected reliability of methods based upon statistical predictions.

A theoretical model for the functioning of expert judgement

In order to assess the reliability of a standard maintaining approach based upon expert judgement it is first necessary to determine the extent to which examiners are able to distinguish between the underlying quality of different examination scripts. Furthermore, it is also necessary to better understand the relationship between the marks awarded to a given script and its likely level of holistic quality.

The first question can be neatly captured by the rank-ordered logit model (Allison and Christakis, 1994) which is a modified version of the Bradley-Terry Model (Bradley and Terry(1952)). This model assumes that if we denote the underlying (judge-independent) quality of the ith script by M_i and the *judged* quality of the same script by judge k as M_{ik} then:

$$M_{ik} = M_i + E_{ik} \quad (1)$$

Where all the E_{ik} are independent and identically distributed according to a standard Gumbel distribution (see Gumbel (1954)). Since the variance of a standard Gumbel distribution is fixed, the above equation implies that the reliability with which judges are able to distinguish higher from lower quality scripts will be determined by the extent to which the M_i in the above equation are spread out.

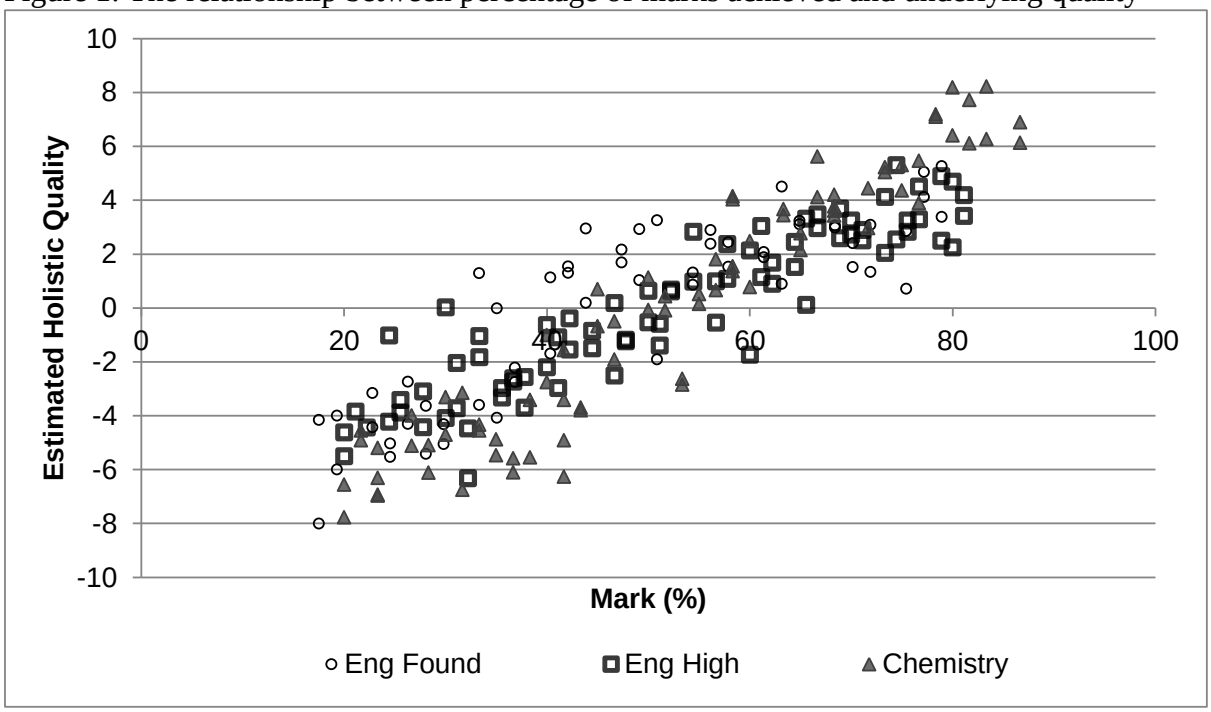
Given a set of data from a comparative judgement exercise (such as a set of rankings given to different scripts by different judges) it is possible to estimate the underlying quality (the M_i) of the scripts included in the exercise using a maximum likelihood approach. Once this is done, it is then possible to examine the extent to which underlying quality (such as is explored in a comparative judgement exercise) relates to the number (or rather the percentage) of marks awarded to different scripts. This enables us to meet the second requirement for estimating the reliability of using such a procedure to maintain standards; to understand the relationship between the marks awarded to a given script and its likely level of underlying quality.

Statistical analysis was undertaken to explore the relationship between the marks awarded a script and its underlying quality for data from three existing rank-ordering studies, chosen purely on the basis of the easy availability of the data. Two of the studies were based on rank ordering of foundation and higher tier English GCSE units and are described more fully in Gill, Bramley and Black (2007). The third set of data was from a rank ordering study of a higher tier Chemistry GCSE unit and is described more fully in Bramley (2009). The three units used in the investigation had 57, 90 and 60 marks available respectively (for English Foundation Tier, English Higher Tier and Chemistry).

For each script in each rank-ordering study, a measure of the holistic quality was estimated using the rank-ordered logit model. The relationship between these estimates of underlying quality and the percentage of marks achieved on each script is shown in figure 1. This shows a clear relationship between the number of marks achieved and the estimated of underlying quality of a script. However, what is more remarkable is the degree of similarity in the

relationships across the three different examinations included in the three separate rank ordering studies.

Figure 1: The relationship between percentage of marks achieved and underlying quality



Using the data in figure 1, a regression model was estimated to capture the relationship between the percentage of marks awarded to a script and its underlying quality[1]. This regression model, combined with the basic assumptions of the rank-ordered logit model, can be used to simulate the likely outcomes of different comparative judgement exercises. This allows us to compare the empirical results from a number of such research studies to the results we would expect given our theoretical model.

Examining previous studies in the light of the theoretical model

Gill and Bramley (2013)

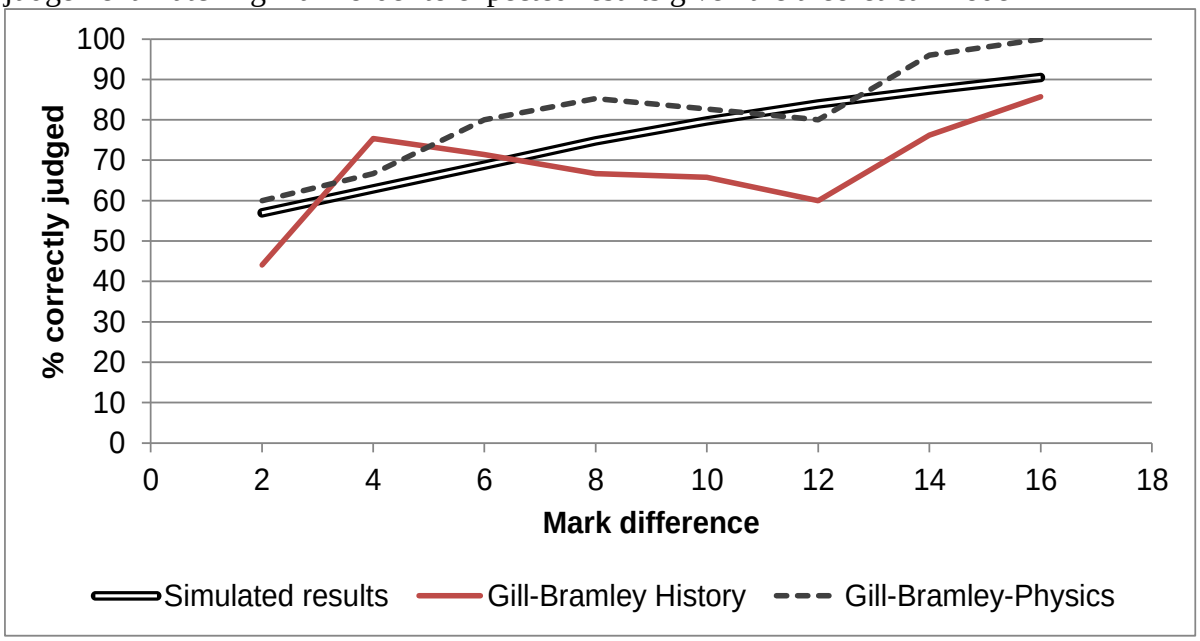
To begin with we compare the empirical results of a research study by Gill and Bramley (2013) to results that would be expected from our theoretical model. In particular, their research examined the chances of individual experts' judgements of which of two scripts is better coinciding with which of the two scripts was awarded a greater number of marks. Using two 90 mark A level examinations in History and Physics, this probability was examined for scripts that differed by between 2 and 16 marks.

Our own estimates of the probabilities of interest were calculated (via simulation) using the theoretical model described in the previous section. The results of these calculations are shown in figure 1 alongside the empirical results found in the original research. Note that, in general, the empirical results were based upon 15 judgements by each of 5 judges at each mark difference. This means that at each mark we have 75 observations. If these were independent observations a 95% confidence interval for each of these empirical estimates would be roughly plus or minus 10 percentage points. With this in mind, it can be seen that

[1] Specifically that $M_i=0.18*(\text{percentage of marks achieved})+\varepsilon_i$ where $\varepsilon_i \sim N(0,1.2)$.

the expected results given in our theoretical model closely match the results from this independent empirical research study.

Figure 1: Comparing results published in Gill and Bramley (2013) on the chances of relative judgement matching mark order to expected results given the theoretical model



Baird and Dhillon (2005)

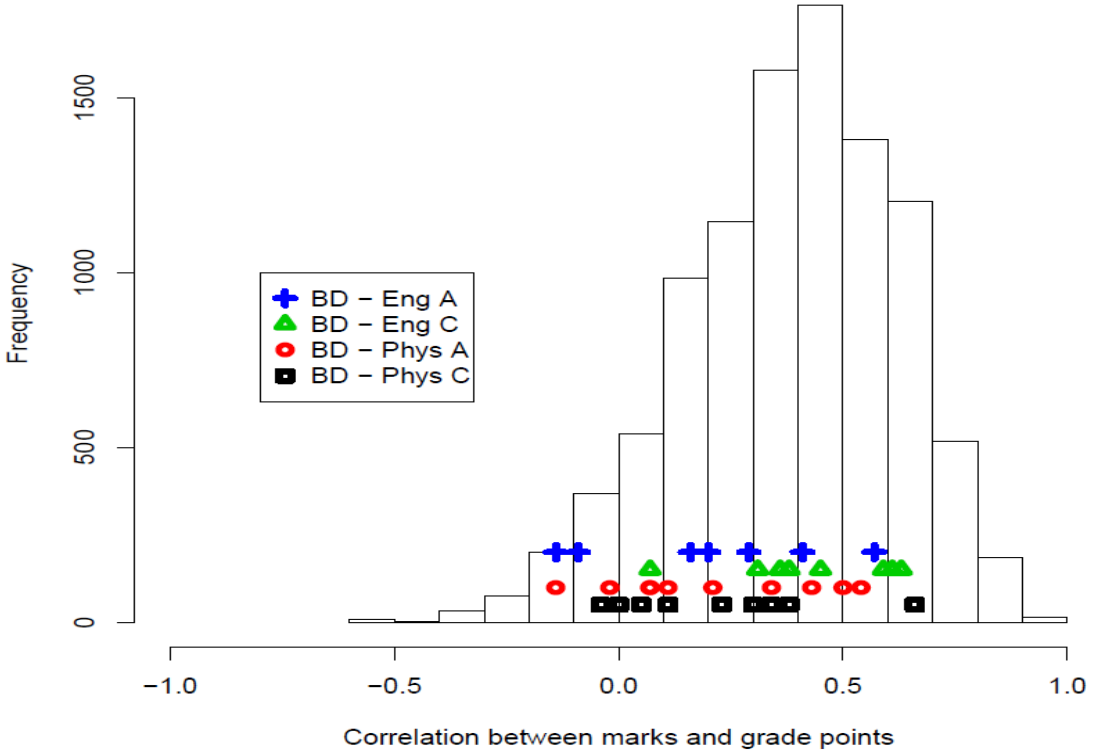
This study contains a section entitled "Can examiners successfully distinguish grade-worthiness within a small range of marks?" which explores comparative judgement for two examination papers (from an English GCSE and a Physics A level) each of 50 marks. Within each of these papers, for both the A grade and C grade boundaries, 14 scripts were identified; two scripts at each mark between 3 marks below and 3 marks above the grade boundary. These scripts were then sent to a number of expert judges (8 for English, 10 for Physics) who were asked to place the scripts into 7 "grade points" based upon their judged underlying holistic quality.

Results presented within the report included the correlations between grade points assigned by expert judgement and the marks awarded to a script. These correlations were calculated for each judge separately for each subject, for both the scripts close to the grade A boundary and for the scripts close to the grade C boundary. The values of these correlations are shown by the points marked in figure 2. As can be seen, these correlations are almost all below 0.5 and many of them are close to (or below) zero. These low correlations were taken to indicate a lack of reliability in the type holistic judgement exercise that might be required in the context of grade awarding. Such evidence led the authors to conclude that "the task of distinguishing grade-worthiness of scripts that vary by only a small range of marks is *impossible*, even for highly expert judges" (emphasis added).

However, using our theoretical model, we are able to estimate what we might have expected the results of their study to show. These results are shown by the histogram element of figure 2, which shows the expected distribution of such correlations based upon simulated version of the same study. As can be seen, all of the observed correlations are within the range expected from the mathematical model albeit just a little lower than expected. However, it is

worth remembering that the empirical correlations are by no means independent of one another. A single script with an unusual level of underlying holistic quality given the number of marks, will affect the correlations for all judges as they are all examining the same set of scripts. Indeed one of the most interesting features of figure 2 is the width of the distribution shown by the histogram. That is, simulations based on exactly the same mathematical model can lead to wildly differing results. For example, simulated correlations for a single judge are frequently as high as 0.75 and often as low as zero. This indicates that using such a small number of scripts in analysis (and the same scripts across all judges) has probably led to a lack of reliability in the study itself.

Figure 2: Comparing reported correlations between mark order and rank order from Baird and Dhillon (2005) to expectations from our theoretical model



Thus, we can see that the empirical results generated by Baird and Dhillon are within the range we would expect given our theoretical model. In other words, our analysis shows that, these results are more or less exactly what we would expect within the same theoretical model underpinning research studies that have been used to promote rank-ordering as a method of maintaining standards. The apparently poor results are likely to be a result of the small number of scripts used within the study and the highly restricted mark range. Similar comments can be made for another study by Forster (2005) with similar results to those presented by Baird and Dhillon.

The reliability of judgemental methods to setting grade boundaries compared to statistical approaches

Having developed our mathematical model, and demonstrated that it can predict empirical results presented both by authors who are supportive and by those who are critical of the potential of expert judgement, we shall now use our model to explore just how reliably comparative judgement could be used within grade awarding.

For the purposes of this paper will examine a method based upon *blind comparative direct judgement*, requiring the following steps:

- Assemble a **number of expert examiners**.
- For each grade of interest, assemble a **number of historical benchmark scripts** with marks at the grade boundary.
- Assemble a **number of scripts from the new examination** such that their marks are evenly spread across a 21 percentage mark range from 10 percentage points below an expected grade boundary (e.g. the grade boundary from the previous year) to 10 percentage points above this expected boundary.
- For each examiner:
    - A benchmark script is randomly selected from the available benchmark scripts
    - A script is randomly selected from the available scripts on the new test.
    - The examiner now decides whether the benchmark script or the new script is superior.
    - This process is repeated for 30 pairs of scripts.
- The grade boundary on the new test is estimated to be the mark on the new examination at which examiners have a probability of at least 50% of judging new scripts to be superior to a randomly chosen benchmark script. This point can be identified using logistic regression.

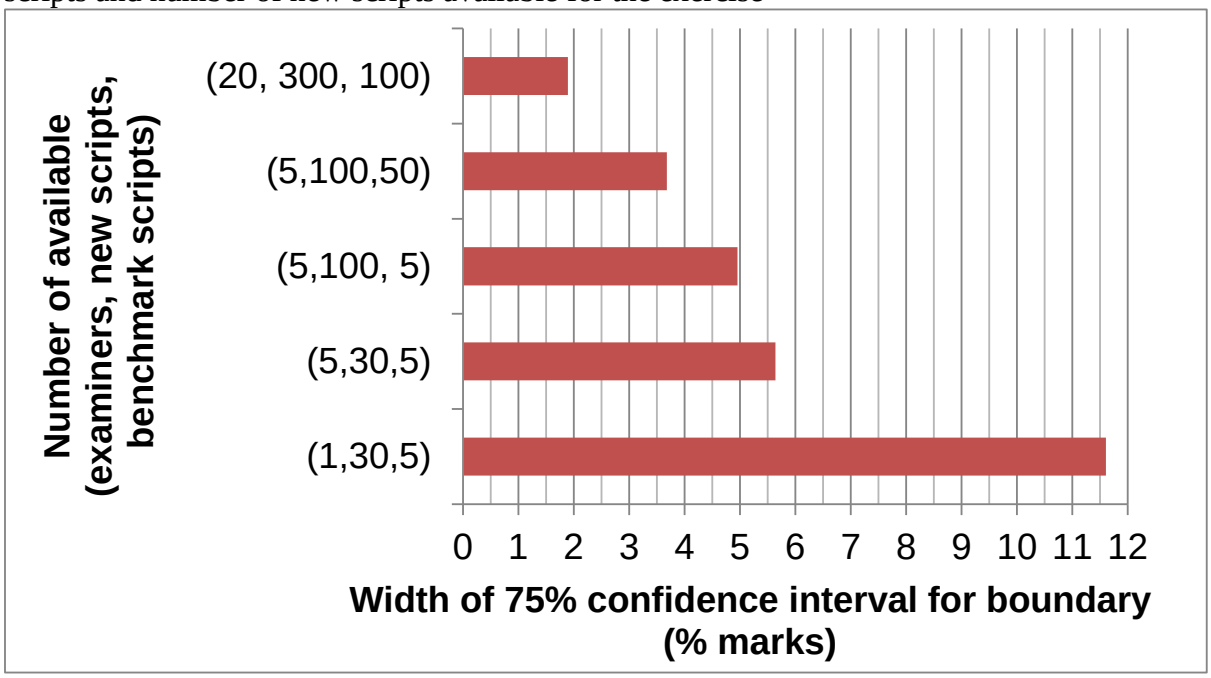
Given the need to sample at random from both benchmark and new scripts, and the potential large scale of the exercise it is possible that this process would need to be conducted on-screen rather than being paper-based.

Making use of the mathematical model described earlier, a simulation study was undertaken to establish the likely stability of grade boundaries derived using this method. A number of factors (highlighted in bold) in the above description have been left unspecified and the simulations also explored how the reliability of the method might change for different values of these parameters (number of examiners, number of benchmark scripts, and number of new scripts). For each combination of parameters examined, 1000 simulations were used to calculate the expected width of a 75% confidence interval for the grade boundary. A subset of the results is shown in figure 3.

The results in figure 3 show that all 3 of the above parameters are crucial to increasing the reliability of the method. Involving larger numbers of examiners and sampling from larger numbers of scripts (both benchmark and on the new test) tends to increase stability. Indeed, whilst a small scale exercise involving only five examiners leads to a 75% confidence interval roughly 5.5 percentage marks wide[2], for the largest scale exercise the width of the confidence interval becomes less than 2 percentage marks wide.

[2] A scenario involving only a single examiner is also shown for the sake of interest but would not be considered as a serious approach in practice.

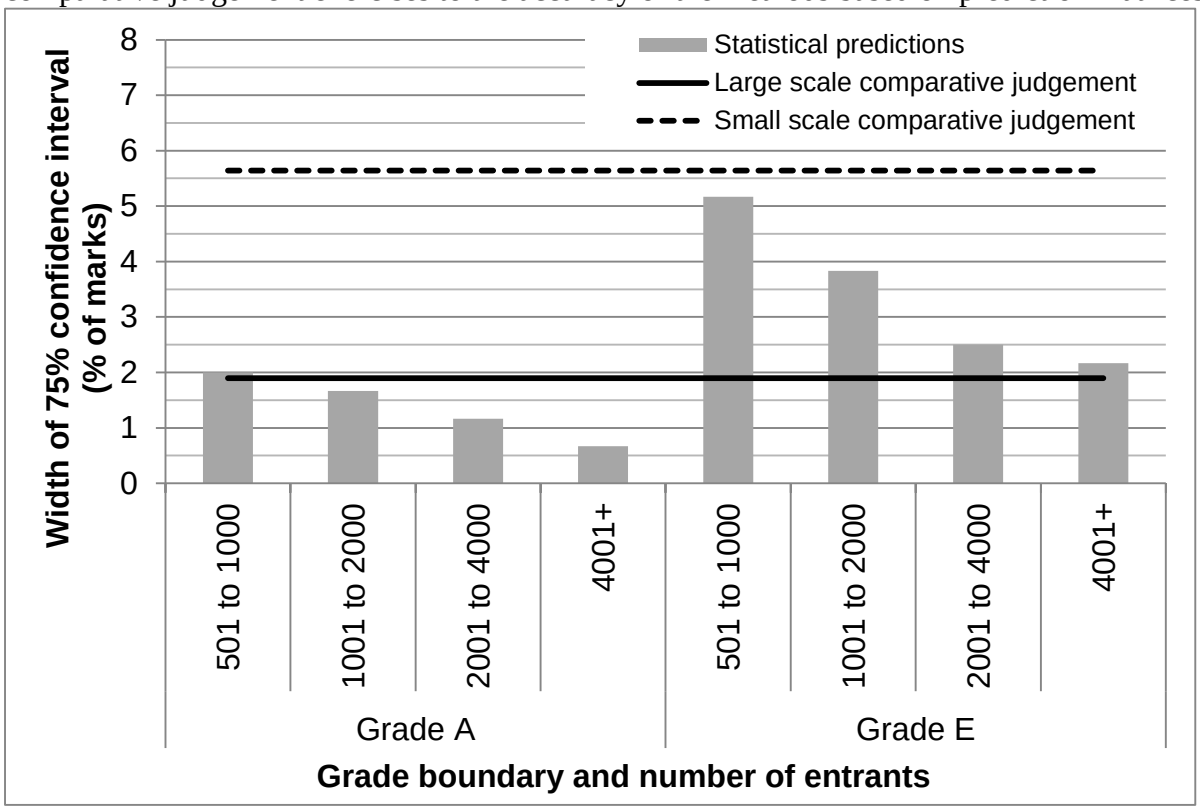
Figure 3: Expected 75% confidence intervals for grade boundaries set by direct blind comparative judgement method dependent upon number of examiners, number of benchmark scripts and number of new scripts available for the exercise



Finally we can compare the results in figure 3 to estimates of the likely reliability of a method based upon statistical predictions. Specifically we compare the 75% confidence intervals from one of the smallest (5 examiners, making 30 paired comparisons sampled from an available 5 benchmark scripts and 30 new scripts) and the largest (20 examiners, making 30 paired comparisons sampled from 100 benchmark scripts and 300 new scripts) comparative judgement exercises to the 75% confidence intervals expected from statistical predictions. To derive the confidence intervals for statistical predictions we make use of the work of Benton and Lin (2011). Their calculations yield the expected size of confidence intervals for the *percentage of students* predicted to achieve each A level grade dependent upon their prior attainment (at GCSE) for different samples sizes of students. These results are combined with a typical score distribution from a large scale OCR A level to produce confidence intervals in terms of the *percentage mark range* on the test.

The reliability of the statistical approach and the expert judgement approach are compared for different available sample sizes in figure 4. This shows that comparative judgement is unlikely to provide similar reliability to statistical predictions, if it is based upon a small scale exercise, except at grade E and if the statistical predictions are based upon small samples. However, if it is a large scale exercise, comparative judgement can sometimes be just as reliable as a statistical approach at grade A and superior at grade E.

Figure 4: Comparing expected accuracy of large scale and small scale direct blind comparative judgement exercises to the accuracy of the methods based on prediction matrices



Conclusion

The results of several previous studies examining the accuracy of examiners' comparative judgements all fit within the same theoretical framework. This includes studies that are supportive of the idea of using expert judgement (such as Gill and Bramley, 2013) and studies that are critical of its potential (such as Baird and Dhillon, 2005, or Forster, 2005).

Whilst it would be foolish to expect that the theoretical model used in this paper would apply universally, it is striking to note how well it works across a number of research projects with different objectives. This suggests that the results from a number of different research studies are far more compatible than has previously been recognised.

Simulation studies suggest that, under certain circumstances, such as where we have relatively small numbers of entrants to an examination, methods based on comparative judgement could potentially provide a more reliable way of setting grade boundaries than using statistical predictions. Having said this, research examining how the method functions in practice would be required before any such statement could be made with certainty.

References

Allison, P.D., and Christakis, N.A. (1994). Logit models for sets of ranked items. *Sociological methodology*, 24, 199-228.

Baird, J., and Dhillon, D. (2005). *Qualitative expert judgements on examination standards: Valid, but inexact*. AQA research report RPA_05_JB_RP_077. Guildford: AQA.

Benton, T., and Lin, Y. (2011) *Investigating the relationship between A level results and prior attainment at GCSE*. Coventry: Ofqual.

Bradley, R., and Terry, M. (1952) The rank analysis of incomplete block designs: I: The method of paired comparisons. *Biometrika*, 39, 324-345.

Bramley, T. (2009). *The effect of manipulating features of examinees' scripts on their perceived quality*. Paper presented at the Association for Educational Assessment – Europe (AEA-Europe) annual conference, Malta.

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

Gill, T., Bramley, T., and Black, B. (2007). *An investigation of standard maintaining in GCSE English using a rank-ordering method*. Paper presented at the British Educational Research Association annual conference, London, UK.

Gill, T., and Bramley, T. (2013). How accurate are examiners' holistic judgements of script quality?. *Assessment in Education: Principles, Policy and Practice*, (ahead-of-print), 1-17.

Gumbel, E.J. (1954) *Statistical theory of extreme values and some practical applications*. Applied Mathematics Series, 33. U.S. Department of Commerce, National Bureau of Standards.

Novakovic, N., and Suto, I. (2010) The reliabilities of three potential methods of capturing expert judgement in determining grade boundaries. *Research Matters: A Cambridge Assessment Publication*, 9, 19-24.

Stringer, N. S. (2012). Setting and maintaining GCSE and GCE grading standards: the case for contextualised cohort-referencing. *Research Papers in Education*, 27(5), 535-554.