Greatorex, J., Hamnett, L. & Bell, J.F. (2003). *A comparability study in GCE Chemistry including the Scottish Advanced Higher Grade*. A study based on the Summer 2002 examination and organised by the Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for OCR on behalf of the Joint Council for General Qualifications.

Greatorex, J., Johnson, C. & Frame, K. (2001). Making the grade – developing grade profiles for accounting using a discriminator model of performance. *Westminster Studies in Education*, **24**, 2, 167–181.

Greatorex, J., Novaković, N. & Suto, I. (2008). *What attracts judges' attention? A comparison of three grading methods*. A paper presented at the International Association for Educational Assessment Conference, Cambridge, September 2008.

Guthrie, K. (2003). *A Comparability Study in GCE Business Studies and VCE Business, A review of the examination requirements and a report on the cross moderation exercise*. A study based on the Summer 2002 Examination and organised by the Edexcel on behalf of the Joint Council for General Qualifications.

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

Laming, D. (2004). *Human judgment: The eye of the beholder*. London: Thomson.

Murphy, R., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmut, J. & Gower, R. (1995). *The Dynamics of GCSE Awarding*. Report of a project conducted for the School Curriculum and Assessment Authority. School of Education, University of Nottingham.

OCR (2004). OCR AS GCE Business Studies (3811) OCR Advanced GCE in Business Studies (7811) Approved Specification, Revised Edition. OCR. [online.] Available at: http://www.ocr.org.uk/qualifications/as_alevelgce/business_studies/documents.html

Pollitt, A. & Elliott, G. (2003a). Monitoring and investigating comparability: a proper role for human judgement. Qualifications and Curriculum Authority Seminar on 'Standards and Comparability', UCLES, 4th April 2003.

Pollitt, A. & Elliott, G. (2003b). Finding a proper role for human judgement in the examination system. Qualifications and Curriculum Authority Seminar on 'Standards and Comparability', UCLES 4th April 2003.

Qualifications and Curriculum Authority (2008). GCSE, GCE, and AEA code of practice 2008. QCA: London.

Scharaschkin, A. & Baird, J. (2000). The effects of consistency of performance on A Level examiners' judgements of standards. *British Educational Research Journal*, **26**, 3, 343–357.

Townley, C. (2007). *Australian Education Systems Officials Committee – Secondary Schools Reporting – A study to examine the feasibility of a common scale for reporting all senior secondary subject results*. Victorian Curriculum and Assessment Authority: Melbourne.

'Key discriminators' and the use of item level data in awarding

Tom Bramley Research Division

Introduction

As more examination papers in general qualifications (GCSEs and A levels) are scanned and marked on screen, the marks on individual questions or question parts are collected automatically, and are referred to as item level data (ILD). The analysis of ILD is available for use in awarding meetings (where the grade boundaries are decided). This article discusses the theoretical rationale for using ILD in awarding, presents some possible formats for displaying data, and suggests ways in which the data could be used in practice.

For many examinations (whether marked on screen or not), the Principal Examiner (PE) will have produced a list of the questions which they expected to be 'key discriminators' at particular grade boundaries. This information might come from the test blueprint (for example, if each question on a test was 'targeted' at pupils at a particular grade or level), or it might come from the PE's (and their marking team's) experience of marking the papers – for example, if during the course of marking the paper they noticed which questions seemed to be discriminating well at particular grades or levels.

The (often unspoken) assumption behind identifying these 'key discriminators' is that by focussing on performance on these questions when making judgements about scripts in the awarding meeting, the awarding panel will use their time and effort most efficiently and be best able to identify the overall score on the test which represents the same performance standard as the corresponding grade boundary set in previous sessions.

The Guttman pattern – an idealised scenario

Imagine that we have a test consisting of ten dichotomous items (items scored 1 or 0). The scores on such a test fit a Guttman[1] pattern if success on an item implies success on all easier items and failure on an item implies failure on all harder items. If the columns represent the items with the easiest item at the left and the hardest item at the right, and the rows represent examinees with the least able at the top and the most able at the bottom, then a Guttman pattern for scores of 23 examinees on this 10-item test might look like Table 1 below.

If the score data fit this idealised pattern then all scripts on the same test total would show exactly the same performance (in terms of which items were answered correctly and incorrectly). In other words, every script perfectly represents the performance of all examinees with the same test score. Furthermore, there is a 'simple order' in the raw scores. Each increasing test total implies that the examinee has achieved

1 Louis Guttman (1916–1987) was an American psychologist. See http://en.wikipedia.org/wiki/Guttman_scale for more information.

Table 1: Illustration of Guttman pattern of test scores

	Easy →									*Hard*	
	Q4	*Q2*	*Q1*	*Q7*	*Q5*	*Q8*	*Q3*	*Q10*	*Q9*	*Q6*	*Total*
E01	0	0	0	0	0	0	0	0	0	0	0
E02	1	0	0	0	0	0	0	0	0	0	1
E03	1	1	0	0	0	0	0	0	0	0	2
E04	1	1	0	0	0	0	0	0	0	0	2
E05	1	1	1	0	0	0	0	0	0	0	3
E06	1	1	1	0	0	0	0	0	0	0	3
E07	1	1	1	0	0	0	0	0	0	0	3
E08	1	1	1	1	0	0	0	0	0	0	4
E09	1	1	1	1	0	0	0	0	0	0	4
E10	1	1	1	1	0	0	0	0	0	0	4
E11	1	1	1	1	0	0	0	0	0	0	4
E12	1	1	1	1	1	0	0	0	0	0	5
E13	1	1	1	1	1	0	0	0	0	0	5
E14	1	1	1	1	1	1	0	0	0	0	6
E15	1	1	1	1	1	1	0	0	0	0	6
E16	1	1	1	1	1	1	0	0	0	0	6
E17	1	1	1	1	1	1	1	0	0	0	7
E18	1	1	1	1	1	1	1	0	0	0	7
E19	1	1	1	1	1	1	1	0	0	0	7
E20	1	1	1	1	1	1	1	1	0	0	8
E21	1	1	1	1	1	1	1	1	0	0	8
E22	1	1	1	1	1	1	1	1	1	0	9
E23	1	1	1	1	1	1	1	1	1	1	10

everything that examinees with a lower test total have achieved, plus one more item correct.

In a situation like this, the task of the award meeting would be to decide on the pattern of performance which was worthy of the particular grade – and this could be done by considering individual items. For example, suppose that in the above scenario the test is a simple pass-fail test, and a total of 6 out of 10 is under consideration for the pass mark. Inspection of Table 1 shows that it is success on Q8 which distinguishes those with a total of 6 out of 10 from those with a total of 5 out of 10. The content of Q8 could therefore form the basis of a discussion as to whether this was indeed an appropriate cut-score.

This could allow genuine criterion referencing in standard setting. If we imagine our example is a functional maths test, and that Q8 involves calculating a percentage, if it was deemed essential that the 'minimally competent examinee' should be able to calculate a percentage, then 6 out of 10 is the lowest score on the test which guarantees this.

Also, once the standard has been set, standard-maintaining in such a scenario is also straightforward. By including a similar (ideally identical) item in a future test we might anchor the new test to the old simply by finding the lowest test total on the new test guaranteeing success on this item, assuming of course that the new test also produces scores in the Guttman pattern. Thus it would not matter if the new test were easier or more difficult than the original test – the cut-score would vary accordingly.

The traditional item analysis statistics of facility (mean item mark as a proportion of maximum item mark) and discrimination are shown below in Table 2.

Table 2: Facility values and discrimination indices for example data in Table 1

	Easy →									*Hard*
	Q4	*Q2*	*Q1*	*Q7*	*Q5*	*Q8*	*Q3*	*Q10*	*Q9*	*Q6*
Facility	0.96	0.91	0.83	0.70	0.52	0.43	0.30	0.17	0.09	0.04
Discrimination	0.34	0.44	0.57	0.68	0.77	0.78	0.72	0.59	0.45	0.34

Facility is the mean mark on each item as a proportion of maximum item mark[2].

Discrimination is the Pearson correlation between item score and total score minus that item.

These statistics do not seem especially useful for setting cut-scores at an awarding meeting. The facility values are sample-dependent, and the discrimination indices are both sample-dependent and facility-value-dependent.

It is more informative to consider the relationship between score on test and score on item. This can be presented graphically in what are known as Item Characteristic Curves (ICCs). Figure 1 shows the ICCs for Q1, Q5 and Q10 in our example.

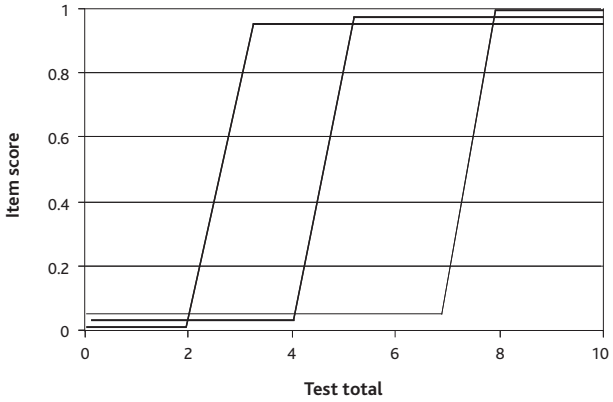


Figure 1: Plot of item score against test total for Q1, Q5 and Q10. (The top and bottom of each ICC should be at 1 and 0, but are separated here for clarity)

These ICCs illustrate the step change in performance on each item with increasing test total score. The slope of the ICC is another indicator of discrimination – in this case it can be clearly seen that each item discriminates very well (perfectly!) by the same amount at a different point on the raw score scale. This information is obscured by using the traditional discrimination statistics (see Table 2).

This kind of display shows the link between performance on the test as a whole and performance on an individual item, and as such is far more relevant to the task which awarders are engaged in when assessing performance on 'key discriminators'.

However, it is virtually impossible in practice to construct tests which produce the deterministic Guttman pattern of responses. This is first because people of the same overall ability differ in their specific knowledge and skills and thus tend to produce different patterns of correct and incorrect responses; and secondly because there are many unknown 'random' variables which might influence a particular response on a particular occasion. In order to overcome both of these factors it would be necessary to use items very widely spaced in difficulty and to administer the test to a population with a very wide distribution of ability. For example, the four-item test below, administered to the entire population of England, might produce a Guttman pattern of responses:

Q1 $2+2=?$

Q2 $^2/_3 \times {}^3/_4 = ?$

Q3 $x^2 - 5x + 6 = 0, x = ?$

Q4 Prove $e^{i\pi} + 1 = 0$

2 For a dichotomous item the facility value is also the proportion of examinees who answered correctly.

But the results of such a test would be extremely uninformative for most educational purposes! Therefore it is necessary to consider another idealisation, but a slightly more realistic one – the Rasch model.

The Rasch model

It is outside the scope of this article to derive or explain the Rasch model – see Wright & Stone (1979) or Bond & Fox (2000) for details. The Rasch model for dichotomous items can be written as:

$$p(x_{ni} = 1 | \beta_n, \delta_i) = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}}$$

where x_{ni} is the score of person n on item i, β_n is the ability[3] of person n, and δ_i is the difficulty of item i.

This model can be considered to be a stochastic form of Guttman's model. This is because the pattern of expected scores from this model is the Guttman pattern. In other words, a person with higher ability has a higher probability of success on every item than a person of lower ability, and every person has a lower probability of success on a more difficult item than on an easier item.

The pattern of observed scores will not exactly conform to the Guttman pattern, but should be a stochastic approximation to it if the data fit the Rasch model. Table 3 shows some simulated data generated to fit the Rasch model, using approximate parameters[4] derived from the data in Table 1.

Table 3 shows that the rank order of examinees has changed slightly, as has the rank order of items, due to the random element in the data generation. More importantly, the data now does not exactly conform to the Guttman pattern, as can be seen by comparing the score patterns of the three examinees with a test total of 5 out of 10. E11's pattern of performance is exactly in line with expectation, but E13 and E14 have both succeeded on some more difficult items and failed some easier items.

Table 3 illustrates the problem of using 'key discriminators' for deciding on cut-scores when data does not fit the ideal Guttman pattern (which it never does). Consider the performance on Q5 by examinees with test scores of 5 and 6. Two of the three examinees with a score of 5 succeeded on this item, whereas only one of the four examinees with a score of 6 did. If consideration of this item were a main focus for awarders then 5 might seem a more appropriate cut-score than 6 – even though examinees with a score of 5 have (by definition) achieved less overall than those with a score of 6, and in particular on other items (e.g. Q8 and Q1) which might not have been deemed 'key discriminators'.

A further issue which is clarified by considering the Guttman pattern in Table 3 is that of examinee (rather than item) score profiles. Some examinees will have an 'unusual' pattern of responses in that they tend to have succeeded on items that they might have been expected to fail, and vice versa. An extreme example in Table 4 is a comparison between examinees E04 and E09 – both with a score of 1. E04 answered the easiest question correctly and failed all the others (as expected), but E09 succeeded on the second hardest question on the test. Given that awarders can only look at a small selection of the scripts on each mark, it would make sense to choose scripts from examinees whose pattern of responses conforms reasonably closely to the Guttman pattern. This is because their responses best exemplify what the test was measuring. In the real world the patterns are far more 'messy' than the neat example in Table 3, which was generated to fit the Rasch model, but the principle is still relevant.

Table 3: Pattern of scores generated by the Rasch model

	Easy →									*Hard*	
	Q2	*Q4*	*Q1*	*Q7*	*Q8*	*Q3*	*Q5*	*Q10*	*Q9*	*Q6*	*Total*
E01	0	0	0	0	0	0	0	0	0	0	0
E04	1	0	0	0	0	0	0	0	0	0	1
E09	0	0	0	0	0	0	0	0	1	0	1
E03	1	1	0	0	0	0	0	0	0	0	2
E05	1	1	1	0	0	0	0	0	0	0	3
E06	0	1	1	1	0	0	0	0	0	0	3
E07	1	1	1	0	0	0	0	0	0	0	3
E02	1	1	1	0	1	0	0	0	0	0	4
E10	1	1	1	1	0	0	0	0	0	0	4
E12	1	1	1	1	0	0	0	0	0	0	4
E11	1	1	1	1	1	0	0	0	0	0	5
E13	1	1	0	1	0	1	1	0	0	0	5
E14	1	1	0	1	1	0	1	0	0	0	5
E08	1	1	0	1	0	1	1	1	0	0	6
E15	1	1	1	1	1	1	0	0	0	0	6
E16	1	1	1	1	1	0	0	0	1	0	6
E17	1	1	1	1	1	1	0	0	0	0	6
E18	1	1	1	1	1	1	1	0	0	0	7
E20	1	1	1	1	1	1	1	0	0	0	7
E19	1	1	1	1	1	0	1	1	1	0	8
E21	1	1	1	1	1	1	1	1	0	0	8
E22	1	1	1	1	1	1	1	1	1	1	10
E23	1	1	1	1	1	1	1	1	1	1	10

3 Ability here does not mean some innate ability or IQ – it simply means the examinee's level on the trait presumed to underlie performance on the test.

4 Note for Rasch experts: it is impossible to estimate parameters for data which exactly fit a Guttman pattern, hence the 'approximate'. The logit difficulties were derived by transforming the facility values, then the abilities were estimated iteratively, arbitrarily assigning reasonable values to scores of 0 and 10 respectively.

Practical application

For an exam with a large entry (say >1000) we can calculate the average score on each item for the set of examinees with each possible score on the test. A table would be one way to present this information for awarders. There should be a general increase in score on each question as test total score increases – this is more likely to be the case for a question which discriminates well (by definition) and the increase is likely to be smoother when the number of examinees is large. At very high and very low test total scores there is likely to be some fluctuation because of the low numbers of examinees.

Table 4 illustrates this kind of information for a GCSE paper, where ILD from approximately 38,000 examinees was captured. Note that the information is shown at the level of the whole question. It would also be possible to show the same information at sub-question level, but such a table would potentially be very large, creating a danger of 'information overload'.

The information in Table 4 might be easier to appreciate if it were presented in graphical form.

The graphs in Figure 2 show one possibility for creating visually informative displays. They simply join the mean y-values (score on question) for each value of x (score on test), and show ± 2 standard errors of each mean. This conveys the information that the location of the mean is more variable at the extremes (or wherever N is low), and also takes into account the spread of the y-values at each value of x (the formula for the standard error is $\sigma/\sqrt{N}$, where s is the standard deviation of the y-values). The individual data points are not shown in the graphs in

Figure 2. This makes the graphs less cluttered, but leaving the points on would emphasise the extent to which there is variability at the individual question level for a given score on the test as a whole.

With a smaller cohort it might be preferable to fit a smoothed line through all the data points, rather than joining up the means in a 'dot-to-dot' fashion. This is an area for further practical experimentation.

Use of item level information in an award meeting

How might the information shown in Table 4 and Figure 2 be used in an awarding meeting? First of all we should note that these whole questions vary quite a lot in terms of maximum marks (Q5 is out of 3 marks and Q9 is out of 12 marks). Q7 was clearly too difficult for most examinees. Q1 and Q8 discriminated best for examinees at the lower end of the score range. The questions with larger mark totals discriminated more smoothly across the score range, as might be expected.

It is possible to identify two approaches for linking information about examinees' performance on individual questions with grading decisions – what we might call a 'prescriptive' approach and a 'maintaining' approach.

On the prescriptive approach, expert judgement combined with grade descriptors might be used to make pronouncements like 'The average borderline grade C examinee ought to score at least 7 marks on Q3'. From the ICC for Q3 or from Table 4, this can be seen to imply a cut-score of around 41. Making several pronouncements of this type on different questions from different topic areas across the paper would produce several potential cut-scores – these could then form a purely judgementally derived range of marks to consider for the grade C boundary. This approach might be more effective at sub-question level because more information could be used (but this would have to be balanced against the dangers of information overload).

On the maintaining approach (which can only work when ILD from two or more sessions are available) the awarding panel would identify questions on the current paper which are similar enough to questions on a previous paper for it to be reasonable to expect performance on them to be equivalent. Now the argument would be along the following lines: 'Last year the borderline grade C examinees (with a test score of 40) averaged 1.2 out of 2 on question 7a, which required them to label a diagram of a cell. This year's question 3b was practically identical, and examinees who averaged 1.2 out of 2 scored 42 on the test overall, suggesting a mark of 42 would be appropriate for this year's boundary.'

Obviously the more 'equivalent' questions that can be identified, the better the linking will be (and this need not always link back to the previous session – links which go back further will help avoid 'drift' in boundaries). There are obviously many caveats which could be raised, such as the extent to which the questions really are equivalent[5], changes over time in topic relevance, drifts in item difficulty, teaching trends etc. – a microcosm of the debates around standards over time more generally! But it is nonetheless a method with some rational justification.

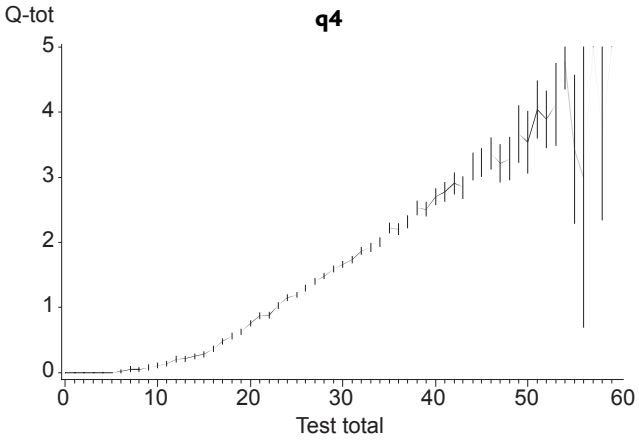
Because of the wide variability in question performance across individual scripts, these judgements based on the ICCs (which show the average performance of the entire examination cohort) might be found to be more effective than judgements based on scrutinising a tiny proportion of the scripts on each mark (as currently happens). It would also prevent individual awarders' judgements being skewed by their impressions from having marked a possibly unrepresentative batch of scripts. For example, using the above data we can see from the graph for Q1 that this question was discriminating most effectively for pupils with a test score of between about 10 and 25 marks. Repeatedly sampling two scripts at random from those with a test total 4 marks apart, and then 1 mark apart gave the results in Table 5 below.

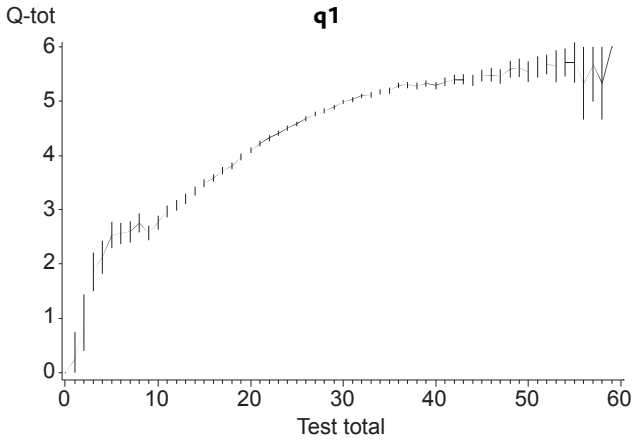
5 It has sometimes been observed that small changes to questions can have a large effect on their facility value, so judgements of equivalence should be made with great care.

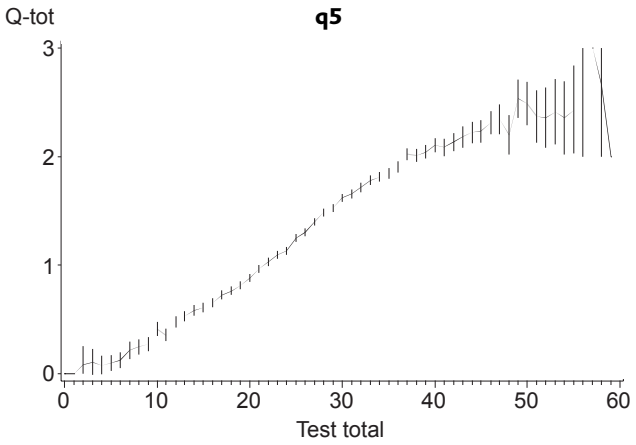
Table 4: GCSE paper – mean scores on each question for pupils with each total test score

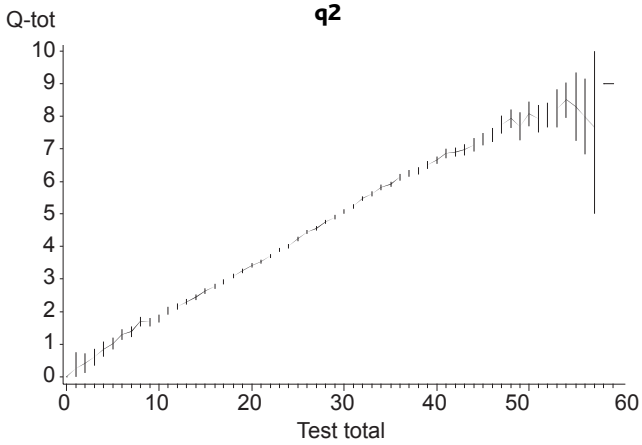
	Max mark→	*6*	*10*	*11*	*5*	*3*	*4*	*4*	*4*	*12*	*7*
Test total	*N*	*q1*	*q2*	*q3*	*q4*	*q5*	*q6*	*q7*	*q8*	*q9*	*q10*
0	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	4	0.25	0.25	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00
2	12	0.92	0.42	0.00	0.00	0.08	0.33	0.00	0.17	0.08	0.00
3	28	1.86	0.61	0.04	0.00	0.11	0.14	0.00	0.14	0.07	0.04
4	39	2.13	0.85	0.13	0.00	0.08	0.15	0.00	0.13	0.36	0.18
5	71	2.54	1.01	0.48	0.00	0.10	0.28	0.00	0.17	0.30	0.13
6	99	2.57	1.29	0.65	0.02	0.12	0.34	0.00	0.31	0.45	0.24
7	117	2.60	1.38	1.03	0.05	0.21	0.49	0.00	0.45	0.53	0.26
8	169	2.76	1.70	1.21	0.05	0.25	0.66	0.01	0.57	0.50	0.30
9	211	2.57	1.68	1.65	0.08	0.27	0.63	0.00	0.82	0.82	0.47
10	263	2.77	1.79	1.83	0.11	0.41	0.79	0.00	1.04	0.76	0.51
11	338	2.97	2.02	1.93	0.14	0.36	0.83	0.01	1.26	0.88	0.62
12	458	3.08	2.15	2.13	0.21	0.48	0.90	0.01	1.29	0.99	0.76
13	579	3.20	2.30	2.32	0.21	0.53	0.88	0.02	1.54	1.19	0.83
14	643	3.35	2.45	2.40	0.25	0.58	1.02	0.01	1.73	1.28	0.94
15	752	3.49	2.63	2.60	0.28	0.61	1.05	0.03	1.82	1.44	1.05
16	925	3.58	2.77	2.68	0.37	0.65	1.12	0.04	2.02	1.63	1.14
17	992	3.72	2.92	2.76	0.48	0.72	1.12	0.03	2.19	1.81	1.25
18	1139	3.81	3.09	2.88	0.56	0.76	1.19	0.04	2.32	2.02	1.33
19	1227	3.97	3.26	2.97	0.63	0.81	1.22	0.04	2.46	2.20	1.44
20	1437	4.09	3.42	3.06	0.76	0.88	1.32	0.06	2.51	2.37	1.53
21	1542	4.22	3.53	3.14	0.87	0.96	1.35	0.06	2.68	2.57	1.61
22	1470	4.32	3.71	3.21	0.88	1.03	1.36	0.06	2.81	2.89	1.73
23	1708	4.41	3.88	3.33	1.03	1.09	1.38	0.08	2.88	3.08	1.83
24	1808	4.50	4.01	3.48	1.15	1.13	1.46	0.08	2.96	3.33	1.90
25	1780	4.58	4.23	3.53	1.19	1.25	1.54	0.09	3.04	3.53	2.02
26	1856	4.68	4.43	3.69	1.29	1.30	1.60	0.10	3.08	3.71	2.12
27	1907	4.77	4.55	3.80	1.40	1.40	1.64	0.11	3.16	3.95	2.22
28	1788	4.82	4.75	3.90	1.49	1.49	1.69	0.11	3.21	4.21	2.33
29	1687	4.89	4.90	4.09	1.59	1.53	1.70	0.12	3.31	4.44	2.44
30	1656	4.99	5.07	4.26	1.66	1.62	1.81	0.13	3.31	4.62	2.54
31	1519	5.03	5.24	4.40	1.73	1.66	1.88	0.13	3.36	4.88	2.69
32	1418	5.10	5.48	4.59	1.86	1.71	1.88	0.16	3.36	5.05	2.81
33	1250	5.12	5.62	4.84	1.92	1.78	1.99	0.12	3.40	5.31	2.89
34	1111	5.17	5.82	5.09	2.00	1.81	2.03	0.15	3.42	5.56	2.95
35	1011	5.19	5.91	5.28	2.21	1.85	2.09	0.15	3.50	5.75	3.06
36	843	5.29	6.13	5.58	2.20	1.90	2.12	0.17	3.49	5.94	3.18
37	766	5.30	6.25	5.68	2.31	2.02	2.24	0.16	3.52	6.20	3.31
38	606	5.29	6.32	6.15	2.53	2.01	2.19	0.18	3.54	6.39	3.41
39	548	5.33	6.51	6.28	2.51	2.04	2.32	0.20	3.52	6.74	3.55
40	497	5.29	6.65	6.58	2.70	2.10	2.42	0.18	3.51	6.95	3.61
41	348	5.36	6.86	6.99	2.77	2.08	2.31	0.20	3.66	7.12	3.65
42	314	5.40	6.90	7.28	2.91	2.13	2.50	0.22	3.53	7.39	3.74
43	236	5.40	6.97	7.60	2.84	2.18	2.54	0.27	3.62	7.59	3.98
44	182	5.38	7.12	7.75	3.16	2.22	2.57	0.27	3.62	7.88	4.03
45	160	5.47	7.30	7.69	3.23	2.23	2.73	0.29	3.71	8.13	4.23
46	136	5.49	7.43	7.99	3.36	2.31	2.78	0.27	3.59	8.43	4.36
47	102	5.46	7.74	8.60	3.22	2.34	2.73	0.48	3.64	8.65	4.16
48	70	5.59	7.93	8.63	3.29	2.20	3.06	0.51	3.66	8.61	4.53
49	45	5.62	7.69	8.47	3.67	2.53	2.91	0.56	3.76	8.89	4.91
50	37	5.54	8.08	8.86	3.54	2.49	3.00	0.89	3.68	9.24	4.68
51	27	5.63	7.93	9.11	4.04	2.37	3.11	0.85	3.67	9.59	4.70
52	28	5.68	8.04	9.43	3.89	2.36	3.00	0.71	3.71	10.04	5.14
53	17	5.65	8.24	9.65	4.12	2.41	3.41	0.88	3.94	9.65	5.06
54	14	5.71	8.50	9.71	4.79	2.36	2.93	0.86	3.86	10.00	5.29
55	7	5.71	8.29	10.43	3.43	2.43	3.29	1.43	3.86	10.29	5.86
56	3	5.33	8.00	10.33	3.00	2.67	4.00	1.33	4.00	11.33	6.00
57	3	5.67	7.67	9.67	5.00	3.00	2.67	1.67	4.00	10.67	7.00
58	3	5.33	9.00	10.00	3.67	2.67	4.00	1.67	3.67	11.33	6.67
59	1	6.00	9.00	10.00	5.00	2.00	4.00	1.00	4.00	12.00	6.00

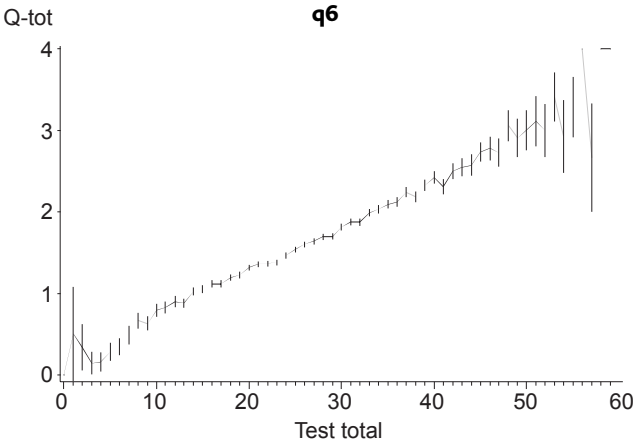
Figure 2: Question score v test score for each question on the GCSE paper

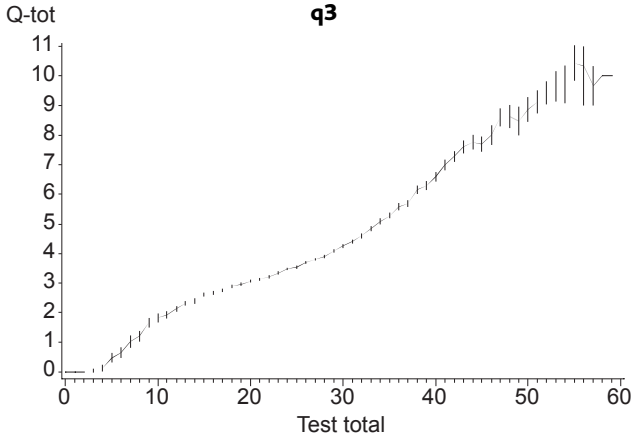


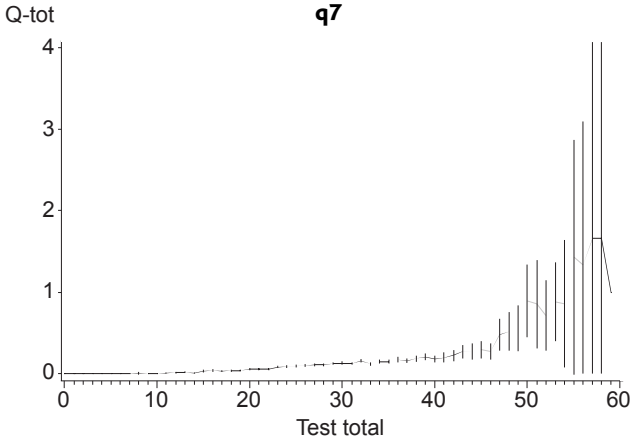


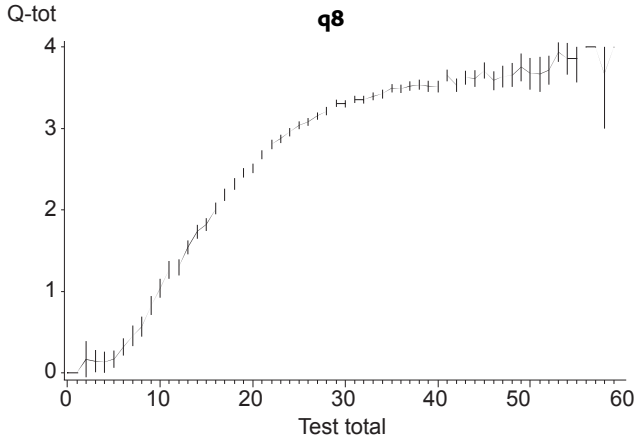


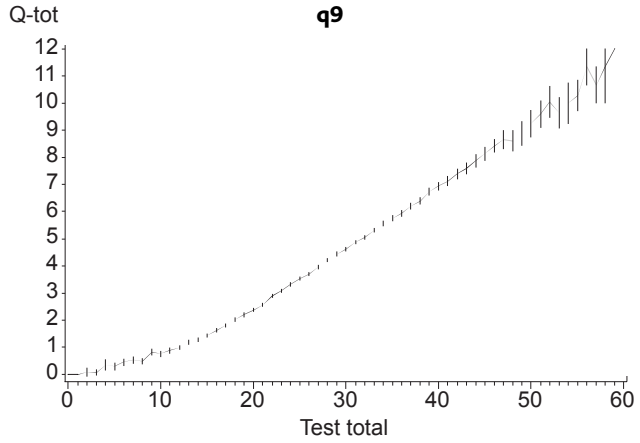












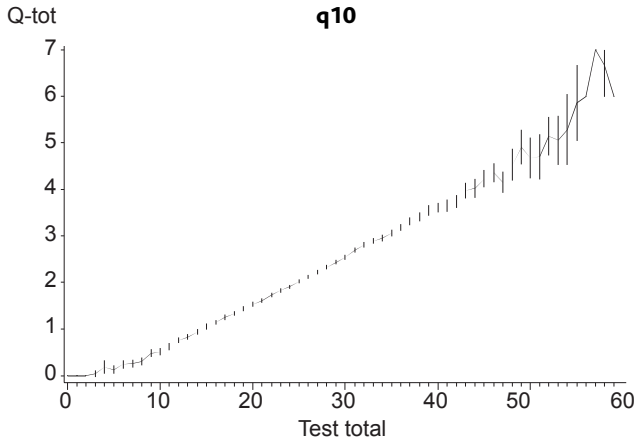


Table 5 shows that even in the most discriminating part of the range, with scripts 4 marks apart on total score the script with the higher test total only scored a higher mark on Q1 about half the time, dropping to around 40% of the time with scripts 1 mark apart. This suggests that script scrutiny might not be a good way to relate performance on 'key discriminators' to total test score.

This is not to suggest that the use of ICC graphs means that script scrutiny can be dispensed with altogether. It does, however, suggest a different focus for the script scrutiny. The ICC graphs can be used to identify the 'key discriminators' for a particular boundary and to derive expectations about the likely range of test total marks corresponding to that boundary, either using the 'prescriptive' or the 'maintaining'

Table 5: Result of 10,000 comparisons of pairs of scripts sampled at random

Performance on Q1	*Test totals 4 marks apart (18 and 22)*	*Test totals 1 mark apart (19 and 20)*
Script with higher test total better	51%	39%
Scores equal	26%	29%
Script with lower test total better	23%	22%

approach described above. The role of the script scrutiny would then be to make a global, holistic judgement about examinee performance on scripts in that range, taking into account performance on all the questions. We must not forget that examinees can compensate for poor performance on the key discriminators by good performance elsewhere. The judgemental task could now perhaps be phrased along the lines 'Would you be happy for scripts in this mark range to receive a C grade?'

In summary, the approach might work as follows:

1. The awarding panel decides for which (if any) questions a 'prescriptive' approach is appropriate and for which questions (if any) a 'maintaining' approach is appropriate.
2. For the 'prescriptive' questions, the awarding panel decides what the minimum mean mark on those questions for examinees at a particular grade boundary should be, using expert judgement and (if appropriate) grade descriptors. The test total mark corresponding to this mark is then located using the ICCs.
3. For the 'maintaining' questions, the awarding panel uses the ICCs to locate the test total mark corresponding to the same question mean mark as that obtained by borderline examinees in a previous session.
4. Steps 2 and 3 should now have created a range of test total marks for consideration at each judgemental boundary. Each range can now be compared with the range produced at the pre-award based on statistical information about score distribution and cohort composition. Hopefully, there will be some overlap between these ranges!
5. The awarding panel can scrutinise scripts in the overlapping range to ratify a particular mark, or narrower range, as appropriate for the boundary in question.
6. The final boundary mark is agreed by the usual process of considering all available evidence.

Implementing this kind of process would create a system where the judgements about scripts can be less influenced by information about pass rates and cohort composition. It has been argued elsewhere (e.g. Black & Bramley, 2008) that this would be desirable.

Conclusion

In summary, consideration of the idealised Guttman pattern of examinee scores on test items leads to the following conclusions:

- If only a small number of scripts is chosen for scrutiny at an award meeting, it is possible that performance on an item designated as a 'key discriminator' will not correspond well with the total score.
- Traditional item analysis statistics (facility values and discriminations) may not be particularly useful for identifying the 'key discriminators' at each grade boundary, but empirical Item

Characteristic Curves (ICCs), ideally based on points plotted at each possible score on the test (when enough data are available), could be much more useful.

- Scripts for the award meeting could be screened to eliminate 'misfitting' examinees with unusual response patterns, or positively selected to aim for responses which conform as well as possible to the Guttman pattern.
- An explicit rationale should be provided for how the item level data will be used in making decisions about grade boundaries – for example, the 'prescriptive' and 'maintaining' rationales described in this article.

References

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

Bond, T. & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B. D. & Stone, M. (1979). *Best Test Design*. Chicago: MESA Press.

Statistics Reports Series

The Statistics Team Research Division

The ongoing 'Statistics Reports Series' provides statistical summaries of various aspects of the English examination system such as trends in pupil attainment, qualifications choice, combinations of subjects and subject provision at school. These reports, produced using national-level examination data, are available in .pdf format on the Cambridge Assessment website: http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Statistical_Reports

In 2009, the following reports were produced:

- Statistics Report Series No.8: Uptake of GCSE AS level subjects in England, 2001–2007
- Statistics Report Series No.9: Numbers achieving 3 A grades in specific A-level combinations by school type and LEA
- Statistics Report Series No.10: Some issues on the uptake of Modern Foreign Languages at GCSE
- Statistics Report Series No.11: Uptake of GCSE and A-level subjects in England by Ethnic Group, 2007
- Statistics Report Series No.12: A-level uptake and results by gender, 2002–2007
- Statistics Report Series No.13: GCSE uptake and results by gender, 2002–2007

Other statistical reports also available on the Cambridge Assessment website are:

- Statistics Report Series No.1: Provision of GCE A-level subjects
- Statistics Report Series No.2: Provision of GCSE subjects
- Statistics Report Series No.3: Uptake of GCE A-level subjects in England, 2001–2005
- Statistics Report Series No.4: Uptake of GCSE subjects, 2000–2006
- Statistics Report Series No.5: Uptake of GCE A-level subjects in England, 2006
- Statistics Report Series No.6: Numbers of A-level examinations taken by candidates in England 2006 and the percentages attaining 3 or more A grades
- Statistics Report Series No.7: The relationship between A-level grade and GCSE grade by subject

Factsheets

The Statistics Team Research Division

In order to make our research accessible to a wider audience we have produced a series of easy-to-read factsheets. The objective of these factsheets is to 'headline' the main findings of some research projects.

They are available in .pdf format on the Cambridge Assessment website:http://www.cambridgeassessment.org.uk/ca/Our_Services/Research

The full research reports can be found in the 'Conference Papers' section of the same website.

As of December 2009, factsheets on the following subjects have been produced:

- AS and A-level choice: Ten factsheets based on the project entitled 'A-level subject choice in England: Patterns of uptake and factors affecting subject references'.
- Emotional Intelligence: Three factsheets based on the project entitled 'Can trait Emotional Intelligence predict differences in attainment and progress in secondary school?'