



CAMBRIDGE ASSESSMENT

Mark scheme features associated with different levels of marker agreement

Tom Bramley

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

Research Division
Cambridge Assessment
1, Regent Street
Cambridge
U.K.
CB2 1GG

Bramley.t@cambridgeassessment.org.uk

www.cambridgeassessment.org.uk

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

Abstract

This study made use of a large database of marker agreement data collected at sub-question (item) level in June 2006 (OCR) and November 2006 (CIE) from a wide range of subjects in public examinations produced by the Cambridge Assessment group¹ taken at ages 16-18 (GCSE, O-Level, IGCSE, AS, A level). The main purpose of the study was to code salient features of the items and their mark schemes and to investigate the relationship of the coded features with the level of marker agreement.

The marker agreement data came from the live processing of the examinations where the marker monitoring is achieved by a hierarchical system where a Team Leader (TL) is responsible for monitoring the quality of the marking by the Assistant Examiners (AEs) in their team. This monitoring is achieved by the TL re-marking a sample of each of their team's allocation of scripts, at one or more points in the marking process. A systematic sample of data from this monitoring process provided TL and AE marks at item (question sub-part) level for around 100 candidates in each of 38 units.

The coding framework was designed to identify relatively coarse, unambiguously identifiable features of the question papers and mark schemes that could apply generally across all subjects and be coded by someone without particular subject expertise. These features included the number of marks available for the item, whether the mark scheme was objective, points-based or levels-based, the amount of space available to the candidate to present their answer, the amount of writing required, the ratio of valid points allowed by the mark scheme to the number of marks available, whether the mark scheme specified qualifications, restrictions or allowable variants to the creditworthy responses, and whether the mark scheme specifically identified wrong answers – i.e. answers not worthy of any credit.

The measure of marker agreement used was the simple P_0 statistic – the proportion of cases with exact agreement between AE and TL. The size and significance of the effect of the coded features on this statistic was assessed using logistic regression modelling.

For both objective items and points-based items up to 10 marks the agreement for an n -mark item closely matched the agreement expected from n independent 1-mark items. After allowing for maximum mark, the amount of constraint in the acceptable responses was strongly related to marker agreement with objective items having about 3 percentage points higher agreement than points-based items worth the same number of marks. The gap widened as the number of marks increased. After allowing for both number of marks and item type, if there were more valid points allowable in the mark scheme than available marks, then this lowered the agreement. These three features alone explained most of the explainable variance in marker agreement (on objective and points-based items), but other coded features were also shown to be significantly related to marker agreement. The findings are discussed in relation to ongoing work at Cambridge Assessment on marking strategies and task complexity.

¹ The Cambridge Assessment Group is a department of the University of Cambridge and a not-for-profit organisation. It comprises three exam boards. Oxford, Cambridge and RSA (OCR) provides qualifications in the UK. University of Cambridge International Examinations (CIE) provides international 14-19 qualifications. University of Cambridge ESOL Examinations (English for Speakers of Other Languages) offers exams for learners of English.

Introduction

Most of the marker agreement analysis reported in research on examinations in England has been at the level of the whole question paper, rather than at the individual item level. The general finding has been that higher correlations among examiners occur on exams containing structured, analytically marked questions than on exams containing essays, and that the less subjective the mark scheme, the greater the reliability (e.g. Murphy, 1978, 1982; Newton, 1996). In a blind multiple-marking study, Massey and Raikes (2006) did concentrate at the item level and found that objective items had higher intra-class correlations (ICCs) than points or levels-based items (see later for a definition of these terms), but that there were differences between subjects in the relation between points and levels-based items, and that the relation of ICC to maximum mark was not always clear-cut. The purpose of the research reported here was to concentrate on agreement at the item level (rather than the candidate level) and to dig deeper into the features of the question papers and mark schemes associated with higher and lower levels of marker agreement.

Recent and ongoing research at Cambridge Assessment¹ is investigating the factors contributing to accurate marking of examinations. These factors can usefully be grouped according to whether they reside in the marker (e.g. factors contributing to marker expertise, such as subject knowledge, level of education, amount of training etc); or whether they reside in the task (e.g. clarity of mark scheme, nature of candidate response, complexity of marking strategy needed etc.). For a brief summary of some of this work, see Suto & Nadas (2007). For a fuller account, see Suto & Nadas (a, in press) and Suto & Nadas (b, in press).

The study reported here is about the second group of factors, that is, those residing in the task. However, the approach taken contrasted somewhat with that of Suto and Nadas. Their work involved only two subjects (GCSE Maths and Physics), a relatively small sample of candidates' work (around 50 scripts per subject), and used more detailed qualitative techniques such as Kelly's Repertory Grid (Kelly, 1955; Jankowicz, 2004) to elicit 'constructs' from Principal Examiners (PEs) which were then related to marking accuracy (at the level of the whole question, not the item). The present study was broader-brush, aiming to identify relatively coarse features of question papers and mark schemes that could apply across a wide range of subjects and be objectively coded by someone without particular subject expertise or examining experience. The much larger size of the data set enabled the relationship between the coded features and marker agreement to be evaluated with more formal statistical modelling than was appropriate in Suto & Nadas (b, in press). The fact that the data were collected at the level of the item rather than the whole question also allowed more secure inferences to be drawn about the effect of the coded features on the agreement statistics. The aims were to discover which features were most strongly related to marker agreement, to discuss any possible implications for question paper (QP) and mark scheme (MS) design, and to relate the findings to the theoretical framework described above.

Data collection

The data came from public examinations taken at age 16 - 18 from (in England) the awarding body OCR, and (internationally) from University of Cambridge International Examinations (CIE). In contrast to the research cited above, this data was collected from the process of marker monitoring in the live examination, as opposed to a research exercise taking place later. The subjects and units involved are shown in Table 1, along with information about the maximum mark for the paper, the time limit, and the number of whole questions and sub-questions (items).

Table 1: Examination papers used in this study.

Exam Board	Unit code	Qualification	Unit name	Tier	Paper total	Time (mins)	# Qs	# items
OCR	1918/01/02	GCSE	Media Studies: Textual Analysis: Moving Image	Both	60	105	4	11
OCR	1937/02	GCSE	History B: (Modern World) Paper 2 (British Depth Study)	n/a	50	90	6	6
OCR	1956/01	GCSE	D&T: Resistant Materials	Foundation	50	60	5	34
OCR	1962/03	GCSE	Mathematics Syllabus A: Paper 3	Intermediate	100	120	23	52
OCR	1983/02 & 1980/02	GCSE	Science: Biology	Higher	100	90	10	52
OCR	1983/04 & 1981/02	GCSE	Science: Chemistry	Higher	100	90	10	50
OCR	1983/06 & 1982/02	GCSE	Science: Physics	Higher	100	90	10	39
OCR	1987/01	GCSE	Geography: Spec B Paper 1	Foundation	90	90	6	68
OCR	2363/01	GCSE	German: Reading	Foundation	50	45	44	44
OCR	2431/01	GCSE	English: Non Fiction, Media and Information	Foundation	63	105	3	5
OCR	2441/02	GCSE	English Literature: Scheme A Unit1 Drama Post -1914	Higher	30	45	12	12
OCR	2520/01	AS	D&T :Product Design 1	n/a	54	72	5	47
OCR	2549/01	A2	Psychology and Crime	n/a	50	90	4	10
OCR	2591/11	A2	History: English and European Themes in History 1763-1996	n/a	120	90	30	30
OCR	2652/01	AS	French: Listening, Reading and Writing 1	n/a	80	90	6	28
OCR	2681/01	AS	Geography A: The Human Environment	n/a	75	60	3	17
OCR	2708/01	AS	English Literature: Poetry and Prose (Open Text)	n/a	60	90	16	32
OCR	2735/01	A2	Media Studies: Media issues and Debates	n/a	90	120	9	9
OCR	2806/01	A2	Biology: Unifying Concepts in Biology	n/a	60	75	5	23
OCR	2816/01	A2	Chemistry: Unifying Concepts	n/a	60	75	5	25
OCR	2822/01	AS	Physics A: Electrons and Photons	n/a	60	60	7	24
OCR	4724/01	A2	Mathematics: Core Mathematics 4	n/a	72	90	9	20
CIE	0452/03	IGCSE	Accounting	n/a	100	105	5	19
CIE	0460/01	IGCSE	Geography	n/a	75	105	6	42
CIE	0470/04	IGCSE	History	n/a	40	60	8	56
CIE	0486/01	IGCSE	English Literature	n/a	60	135	45	45
CIE	0610/06	IGCSE	Biology	1 & 2	40	60	3	15
CIE	1120/01 & 1125/01	O Level	English Language	n/a	60	90	6	12
CIE	4017/01	O Level	Mathematics	n/a	80	120	24	64
CIE	5052/02	O Level	Physics	n/a	80	105	11	51
CIE	5068/02	O Level	Chemistry	n/a	80	105	10	47
CIE	9246/02	A Level	Physics	n/a	90	105	8	47
CIE	9695/03	A Level	English Literature	n/a	50	120	8	16
CIE	9696/01	A Level	Geography	n/a	100	180	11	36
CIE	9700/02	A Level	Biology	n/a	60	75	6	21
CIE	9701/04	A Level	Chemistry	n/a	60	75	5	17
CIE	9708/02	A Level	Economics	n/a	40	90	4	12
CIE	9709/01	A Level	Mathematics	n/a	75	105	10	25

In general, marker monitoring is achieved by a hierarchical system where a Team Leader (TL) is responsible for monitoring the quality of the marking by the Assistant Examiners (AEs) in their team. This monitoring is achieved by the TL re-marking a sample of each of their team's allocation of scripts, at one or more points in the marking process. The procedures that all UK awarding bodies (hence OCR) must follow are specified in the Code of Practice issued by the regulator, QCA (Qualifications and Curriculum Authority). See QCA (2007) for full details. Although not bound by this Code, the procedures followed by CIE are similar. The TL completes a form for each AE, listing the scripts that they have re-marked, the AE's mark and their own mark, at whole paper level.

The aim was to capture data from as many examiners as possible within the subjects sampled. This was done by identifying a sample of examiners whose forms would be collected, then obtaining a sample of the physical scripts from the candidates listed on each form. The guidelines for this sampling are explained below. Once the scripts had been obtained, the marks awarded by AE and TL at item level were keyed into a database.

Sampling of units and scripts

The aim was to cover, between OCR and CIE, two question papers at GCSE level and two at A-level from each of the large-entry subject areas. The analysis reported here used data from 22 OCR units and 16 CIE units. The following sampling plan was agreed to be the most effective use of the resources available for this project:

- Aim for a minimum of 100 scripts from each unit;
- At least 5 scripts from each AE sampled;
- An identical number of scripts from each AE sampled;
- Only use scripts from the last batch of scripts sampled by the TL;
- At least 2 AEs from every team on the panel
- An identical number of AEs sampled from each team;
- Maximise the number of AEs sampled (i.e. more examiners is better than more scripts per examiner, providing the minimum of 5 scripts is not breached)

The final data set contained over 114000 records, with each record containing a mark from an AE and their TL on a single item. (38 units × an average of 100 candidates per unit × an average of 30 items per unit = 114000). This final data set was then checked for a variety of possible errors – the aim being to ensure to as great an extent as possible that any discrepancies in the data between AE and TL were real, and not a result of keying errors or data processing errors.

The coding framework for categorising QP / MS features

The coding framework was developed iteratively – an initial set of features and coding categories was produced after a ‘brainstorming’ discussion with colleagues, and this framework was gradually modified in the light of experience with applying it to some specific QP/MS combinations. In this iterative stage, the main coder and the lead researcher independently applied the coding framework to three of the units. Differences in codings were resolved by discussion, and the framework clarified or amended to reflect the improved understanding. The final version of the coding framework that was used to code all the units is shown in Table 2.

It was clear that many of these features were likely to be related to each other. For example, questions worth more marks will generally require the candidate to generate a response as opposed to select it, give them more space in which to put it, and require more writing. However, one aim of the study was to find which factors were the most relevant in terms of statistical prediction of marking accuracy, so we fully expected that not all of these features would be included in the most parsimonious statistical model.

Table 2: Coding framework used to code different features of the question papers and mark schemes.

QP / MS feature	Valid values	Notes
Maximum mark	1,2, etc.	Use QP/MS to decide what the sub-questions are. Usually square brackets e.g. [2]
Item type	O (objective) P (points-based) L (levels-based)	Use definitions from Massey & Raikes (2006).
Select / generate	S (select) G (generate)	'Select' covers multiple-choice, matching, cloze (where there is a list) etc.
Answer space	N/A '1' up to and including 1 line '2' more than 1 line but less than ½ page '3' ½ page or more	The N/A category is for answers in separate booklets. The 'answer space' does not include the question stem – it is the (maximum) amount of physical space the marker has to scan to locate the answer. This feature can be coded just by looking at the QP.
Writing	N/A '1' one word or simple numerical answer '2' few words / single sentence '3' two or more sentences	The N/A category is for diagrams, sketches, formulas, equations, arrows etc. This feature can be coded by looking at the QP/MS combination.
Apparent marking strategy complexity	S (simple ONLY) C (complex)	Simple = matching or scanning for simple items only . Complex = evaluating, scrutinising and scanning for complex items. Depends on candidate answer, so <i>anticipate the most complex scenario</i> – i.e. err on the side of complex. See extra sheet defining marking strategies ² .
Points to marks ratio	N/A S (same) M (more)	N/A category is for levels-based mark schemes, calculations, QoWC. Same = # correct possible answers equals the number of marks available. More = # correct possible answers exceeds the number of marks available. N.B. Aim to distinguish separate points, not relatively trivial variations in acceptable wording within the same point.
Qualifications, restrictions and variants.	N/A N (No) Y (Yes)	N/A is for levels-based mark schemes. This is to capture where the mark scheme explicitly says (for example) 'allow xxx' or 'also accept yyy' etc; or where a qualification/restriction is given e.g. 'only if...' or 'must also have...'. It also applies to mark schemes where there is 'error carried forward' (ecf).
Wrong answers specified	N/A N (No) Y (Yes)	N/A is for levels-based mark schemes. This is to capture where the mark scheme explicitly specifies an incorrect or unacceptable response, (for example) 'do not accept xxx' or 'NOT yyy' etc.

Hypothesised effects of coding features on marking accuracy

The features to be coded, and the coding categories for each feature, were selected to meet the criteria of being easy to code in a relatively objective way (i.e. not to require specialist subject expertise) and because they were hypothesised to be relevant to marking accuracy, as described below. See Appendix C for some examples of how the coding framework was applied.

Maximum mark [item_max]³

The maximum mark is an easily codable indicator of the length and weight given to the response. We might expect it to be related to the number (or complexity) of cognitive

² This extra sheet is given in Appendix B.

³ The abbreviation for each category given in square brackets is the variable name used in the statistical analysis, which appears in some of the tables and graphs elsewhere in the report.

processing tasks the marker needs to accomplish in marking it. We would probably expect less agreement between markers on questions worth more marks. For pragmatic purposes of predicting marker agreement it may be that this feature captures most of the predictable variation. Massey & Raikes (2006) found that a slightly more sophisticated variant of the maximum mark, the 'implied time restriction' (ITR) which is simply $(\text{item_max} / \text{paper_max}) \times \text{paper_time}$, was the best predictor of the intra-class correlation (an index of agreement among several markers).

Item type [item_type]

This feature was coded using the same definitions of item type as used by Massey & Raikes (2006):

An **Objective** item was here considered to be one where the mark scheme precisely gives the *only* accepted answer (e.g. a single number or word, or a multiple choice item, or an item where a candidate has to rank given information, etc.). Objective items require only very brief, heavily constrained responses from candidates.

A **Points-based** item is one which is marked against a "points" mark scheme. These items generally require brief responses ranging in length from a few words to one or two paragraphs, or a diagram or graph, etc. The key feature is that the salient points of all or most credit-worthy responses may be pre-determined to form a largely prescriptive mark scheme, but one that leaves markers to locate the relevant elements and identify all variations that deserve credit. There is generally a one-to-one correspondence between salient points and marks.

A **Levels** item is one which is marked against a "levels" mark scheme. Often these items require longer answers, ranging from one or two paragraphs to multi-page essays or other extended responses. The mark scheme describes a number of levels of response, each of which is associated with a band of one or more marks. Examiners apply a principle of best fit when deciding the mark for a response.

Massey & Raikes (op. cit.) found that there was more agreement on objective items than on points-based and levels-based items. This coding feature in effect records the amount of constraint in the acceptable answers. We would expect less agreement on the less constrained responses, but then these are often worth more marks (see above) and require more writing (see below) so we might expect these effects to be confounded. Suto & Nadas (b, in press) elicited the constructs of 'Mark scheme flexibility' and 'Single letter answer' from the GCSE Physics PE, and these were shown to be related to marking accuracy (in the expected direction)

Select / Generate [select]

Whereas the item type coding is based on the point of view of the marker, this feature is based on the candidate's perspective. We would expect questions where the candidate selects their response from alternatives presented to them to be more constrained for the marker and hence for it to be easier for markers to agree. Note that it is quite possible for a question where the candidate has to generate their own response to be classified as 'objective' in this coding framework.

Apparent Marking Strategy Complexity [AMSC]

Research into the cognitive strategies used by markers (Suto & Greatorex, 2008) has identified five cognitive strategies involved in marking: 'Matching', 'Scanning', 'Evaluating', 'Scrutinising', and 'No response'. (See Appendix B for further details). Matching and scanning for simple items only are thought to involve 'intuitive' or 'System 1' processing which is, among other things, effortless. Scanning for complex items, evaluating and scrutinising, on the other hand, are thought to involve 'reflective' or 'System 2' processing which is, among other things, effortful. Kahneman and Frederick (2002) describe these two systems in more detail. We might expect items which require System 2 processing to be marked less accurately, and hence for there to be less agreement between AE and TL on such items. This is what was found by Suto & Nadas (b, in press) in GCSE Maths and

Physics. We coded this dichotomously according to whether the apparent strategy would involve System 1 only, or would contain some System 2 processing. However, the difficulty with this feature is that it depends to a larger extent on the actual candidate response than some of the other coded features. We decided to base the coding on the most complex anticipated candidate response – i.e. to err on the side of System 2 when coding this feature.

Answer space [ans_space]

This feature is likely to be strongly related to the maximum mark and the amount of writing required, but it is conceivable that it might have an effect on marker agreement over and above those two features. For example, it might be that the larger the area the marker has to scan visually to locate the correct response, the greater the opportunity for a cognitive processing error, hence lowering the marker agreement.

Writing [writing]

The greater the amount of writing required, the more opportunity there is for candidates to express their answer (correct or incorrect) in a way which is different from what appears on the mark scheme, and thus to require an increasing degree of understanding and interpretation on the part of the marker. We might therefore expect the task of marking questions requiring more writing to be more cognitively demanding, and hence for there to be less marker agreement. Suto & Nadas (b, in press) elicited this construct from the GCSE Physics PE and it was shown to be related to marking accuracy (in the expected direction). For the longer written responses with levels-based mark schemes we might expect differences between the markers in their internalisation of the construct being assessed, and hence differences in marks awarded.

Points to marks ratio [PM_ratio]

We hoped that this feature might be able to distinguish among points-based items worth equal numbers of marks. It seems plausible that where the marker has a wider range of acceptable responses against which to compare the actual responses, the marking task is more complex and we might expect less agreement. As seen in Table 2, this was not always an easy feature to code, because when deciding on the ratio of points to marks the coder has to distinguish between relatively trivial variations in acceptable wording for what is substantively the same point, and substantively different points. Suto & Nadas (b, in press) elicited the similar construct of 'alternative answers' from the GCSE Maths PE and it was shown to be related to marking accuracy (in the expected direction).

Qualifications, restrictions and variants [QRV]

It was difficult to predict what the effect of this feature might be on marker agreement. On the one hand, the purpose of adding qualifications, restrictions and variants to the mark scheme is presumably to clarify to the marker exactly what is worthy of credit. Thus it should make it easier to apply the MS accurately, and therefore items with QRV might have higher levels of agreement. On the other hand, the need to bear in mind all the extra information when considering a response might increase the complexity of the marking task and increase the likelihood of a marker error, decreasing the levels of agreement. It is also possible that these two opposing effects might be different for items with different maximum marks. For example the QRVs might be a help for the larger questions, but a hindrance for the shorter questions. One particular example of a QRV is where the mark scheme allows 'error carried forward' (ecf)⁴. Suto & Nadas (b, in press) elicited this construct from the GCSE Maths PE, and found that questions with ecf were marked less accurately.

Wrong answers specified [wrong]

This is where the mark scheme explicitly mentions a possible response which is *not* worthy of credit. We decided to code this feature separately from the other QRVs because it might be expected in some cases to 'interfere' with the marking strategy. For example, a System 1

⁴ Ecf is where a candidate is not penalised for using an incorrect answer obtained in an earlier part of the question as part of their working for a later part of the question. It is most often seen in questions involving calculations.

strategy of matching text in the answer to text in the mark scheme might result in a marker awarding a mark to a wrong answer which has been explicitly specified on the mark scheme, thus lowering agreement levels. On the other hand, as described above, by clarifying what is not worthy of credit, items with wrong answers specified in the mark scheme might be marked more accurately and hence with higher levels of agreement.

Results

The index of marker agreement chosen was the percentage of exact agreement (P_0) between the AE and the TL. This statistic has the great advantages of simplicity and transparency. It does not indicate the direction of any differences (severity or leniency), but these are arguably of less interest here given that they are likely to pertain to individual markers. Our samples were designed to maximise the number of AEs and TLs sampled within units, but not to obtain large numbers of scripts from individual AEs or TLs. (See Bramley (2007) for a discussion of some of the issues in choosing a suitable statistic to quantify marker agreement).

The P_0 statistic was calculated for each item in each unit, along with a record of how many data points had contributed to the calculation. There were 1091 items in total, but for the purposes of analysis we decided to discount any items for which we had P_0 statistics based on fewer than ten data points. This left a total of 1012 items.

It seemed sensible to compare 'like with like' as much as possible, and to this end we chose to group items by maximum mark. The most natural grouping, based on the numbers of items in the data, is shown in Table 3 below.

Table 3: Distribution of items by maximum mark category.

Max. mark	1	2	3	4	5-9	10-20	21-60	Total
No. of items	329	267	139	87	98	50	42	1012

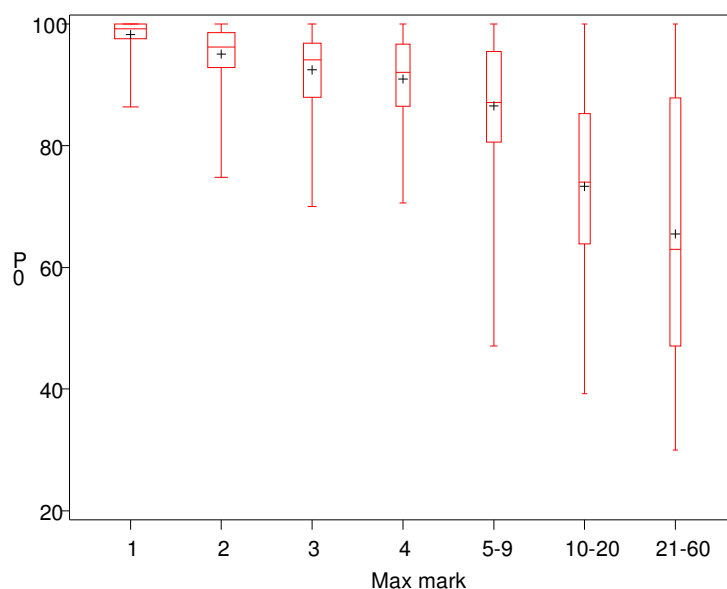


Figure 1: Distribution of P_0 values by item maximum mark. (Width of box is proportional to number of items in each mark category).

We would expect the level of exact agreement between AE and TL to be higher on the lower-mark questions. Figure 1 shows that there was a high level of agreement for the 1-mark items. The median value was around 99% which means that half the 1-mark items had a P_0 value higher than 99%. The vertical length of the box (the interquartile range, IQR) shows

that the middle 50% of the 1-mark items had a P_0 value in the range $\approx 97\%$ to 100% . Figure 1 shows that as the maximum mark increased, the average (median or mean) value of P_0 decreased, and that the spread (IQR) of P_0 values tended to increase.

The following graphs show, for each maximum mark category, the median P_0 value for the items with a given feature coding. Statistical significance was assessed by fitting a logistic regression model because the dependent variable is a proportion (P_0 , the proportion of exact agreement between AE and TL). The general form of all the models described here was:

$$\ln\left(\frac{P_0}{(1-P_0)}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 \quad \text{etc.}$$

where α is the intercept, $\beta_1 \dots \beta_n$ are regression parameters, and $x_1 \dots x_n$ are the independent variables which in this case were maximum mark, item type etc. Table A1 in Appendix A shows which variables were included in each model, and how many items were involved. The parameter estimates from the different regression analyses are given in Tables A2 and A3 in Appendix A. These models are referred to in the text as Model 1, 2 etc. The number of items contributing to each median is shown on the graphs where possible. Appendix D contains full tables of frequencies.

Many of the coded features were only applicable to objective and points-based items. These items tended to be worth 9 marks or fewer. The analysis in Models 1 to 6 was restricted to these items.

Item maximum

The first model (Model 1) just included item maximum mark as a predictor (independent variable). As expected, its parameter was negative (showing that a higher maximum mark was associated with lower agreement), and statistically significantly different from zero. A term for the square of the item maximum was also significant, showing that the relationship was not linear.

Item type

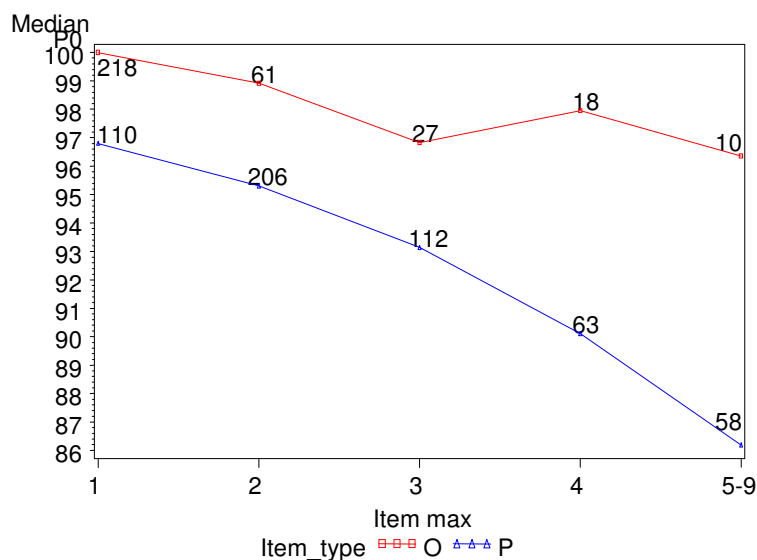


Figure 2: Median P_0 values for objective (O) and points-based (P) items.

Figure 2 clearly shows that for items with a given maximum mark, there was a higher average level of agreement for 'objective' items than for 'points-based' items. The average difference was about 3 percentage points for 1-mark items, growing to about 10 percentage points for 5-9 mark items. In the logistic regression (Model 2) all main effects and interactions were statistically significant. This finding fits the expectation that the amount of constraint in the mark scheme (the essential difference between objective and points-based items) affects the marking accuracy, and agrees with the results of Massey & Raikes (2006).

Points-to-marks ratio (PM_ratio)

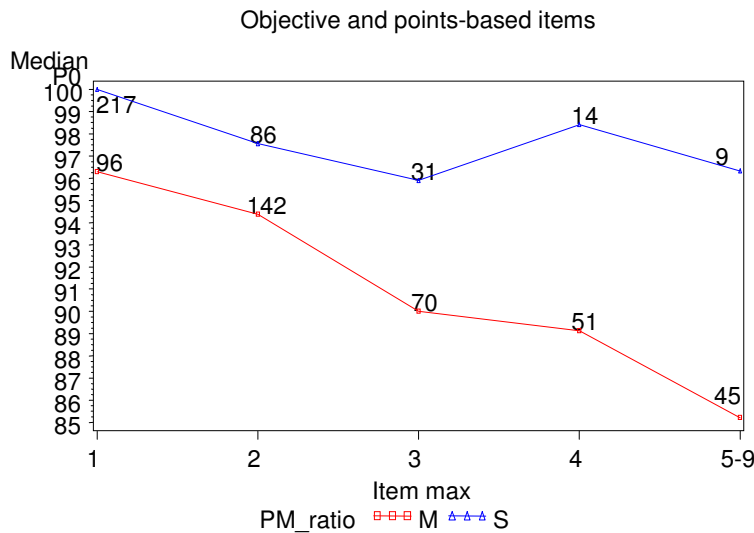


Figure 3: Median P₀ values for objective and points-based items with the same (S) and more (M) points than marks.

Figure 3 shows that for points-based items with a given maximum mark, there was higher agreement for the 'S' items where the number of points equals the number of marks than for the 'M' items where the number of valid points exceeds the number of marks. The differences were around 4 percentage points for 1 and 2 mark items, but larger for the larger items. In the logistic regression (Model 3) all main effects and all 2-way interactions were statistically significant, except for the interaction between points/marks ratio and item type (showing that the effect of the points/marks ratio was not different for the objective and points-based items).

Qualifications, restrictions and variants (QRV)

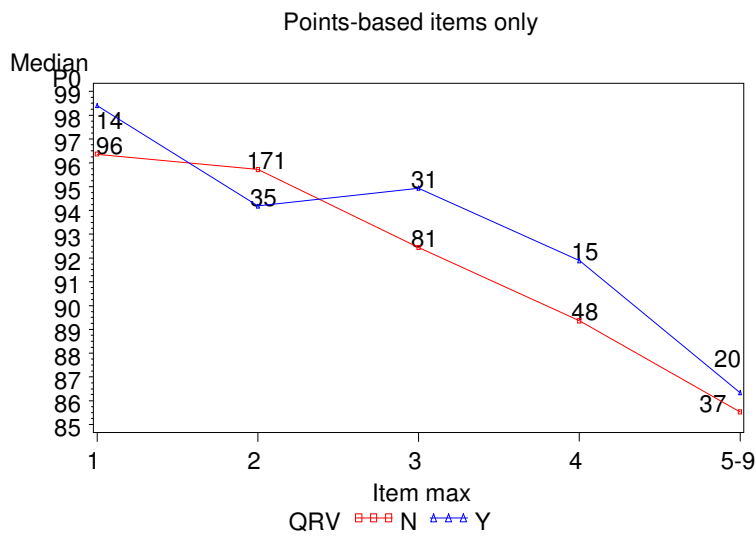


Figure 4: Median P_0 values for points-based items with (Y) and without (N) any QRVs.

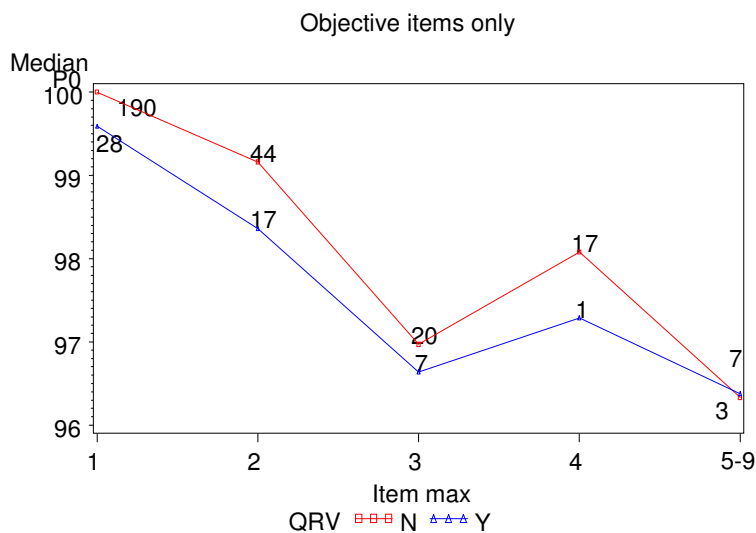


Figure 5: Median P_0 values for objective items with (Y) and without (N) any QRVs.

Figures 4 & 5 show an interesting interaction between item type and the presence of QRVs in the mark scheme. For the points-based items Figure 4, the presence of qualifications, restrictions and variants seemed to increase the level of agreement very slightly. The pattern is spoiled by the 2-mark items, but for the other marks there seemed to be a difference of around 2-3 percentage points. For the objective items, on the other hand, the presence of qualifications, restrictions and variants seemed to reduce the level of agreement, as shown in Figure 5. In the logistic regression (Model 4) all main effects and all 2-way interactions were statistically significant. See the discussion for a possible explanation of this result.

Wrong answer specified (wrong)

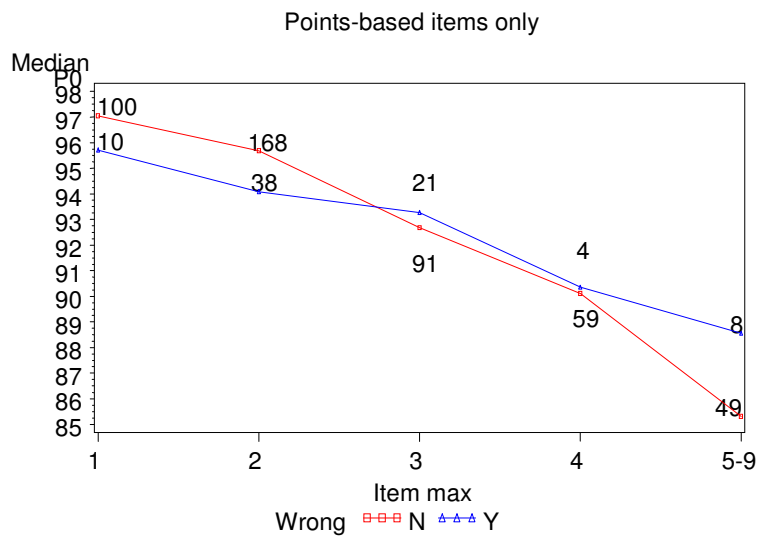


Figure 6: Median P_0 values for points-based items with (Y) and without (N) any wrong answers specified in the mark scheme.

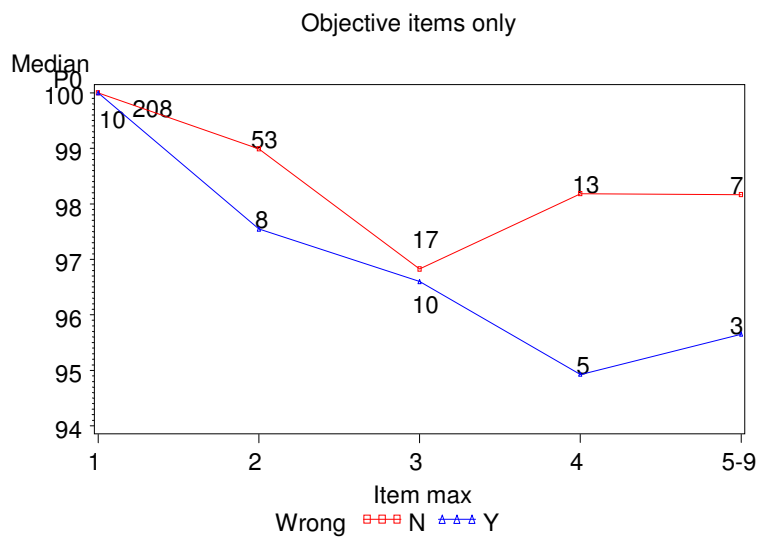


Figure 7: Median P_0 values for objective items with (Y) and without (N) any wrong answers specified in the mark scheme.

As with the QRV, it is interesting to separate the objective items from the points-based items, shown in Figures 6 and 7. The presence of a specific wrong answer in the mark scheme appeared to be associated with lower marker agreement for objective items, and also for the 1 and 2-mark points-based items. In the logistic regression (Model 5) all main effects were statistically significant, as was the interaction between 'wrong' and item type, but not the interaction between 'wrong' and maximum mark, suggesting that chance cannot be ruled out as the explanation for the lower agreement on the 1 and 2-mark points-based items. See the discussion for some possible explanations of this result.

Best predictive model

The logistic regression model which best explained variability in exact agreement (P_0) between AE and TL for objective and points-based items worth up to 9 marks contained main effects for maximum mark, maximum mark squared, item type and points/marks ratio. This result was achieved using both forwards and stepwise selection of variables. There were no significant interactions (see Model 6). The features of 'QRV' and 'Wrong answer specified' did not seem to add any extra predictive capability.

The features of 'answer space' and 'amount of writing required' were applicable to all items (that is, not just objective and points-based items up to 9 marks), although obviously in many places there was little overlap between the different cross-categorisations according to maximum mark and item type, which restricted the scope of the statistical modelling (see Appendix A).

Answer space (*ans_space*)

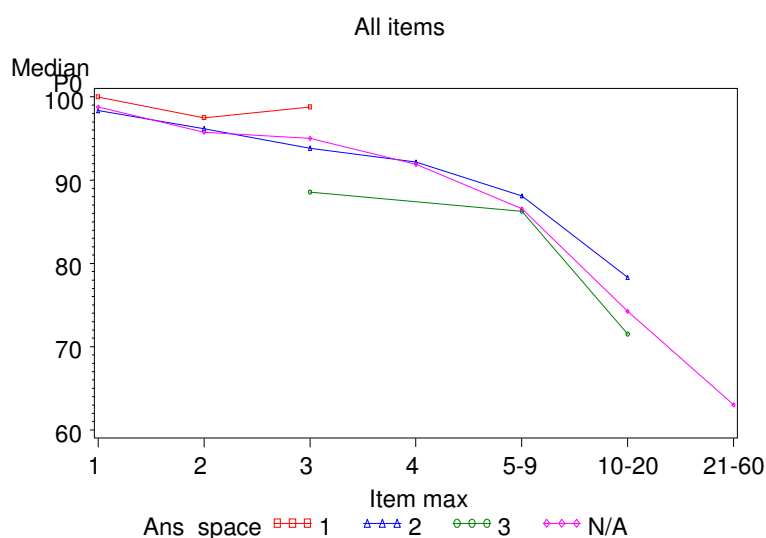


Figure 8: Median P_0 values (all items) for different amounts of answer space.

Figure 8 shows that there was a small effect of the amount of answer space for a given maximum mark, in the expected direction – i.e. slightly higher agreement corresponding to less physical space for the marker to examine to locate the answer. Perhaps the most interesting feature of Figure 8 is the lack of difference between the values for '2' (answer spaces of more than one line but less than half a page) and 'N/A' (the category for responses in a separate answer booklet). This impression was supported by the logistic regression (Model 7). (The levels-based items worth up to 9 marks were included in this analysis, but item type was not included as a predictor in the model). The main effects and interactions relative to the 'N/A' category were significant, except for the effect and interactions for answer space category '2'.

This suggests that although there may be reasons for favouring combined question-answer booklets over separate answer booklets (or vice versa) in terms of the quality and quantity of the candidate's response (Crisp, 2008), the effect on marker agreement is not one of them.

Writing (writing)

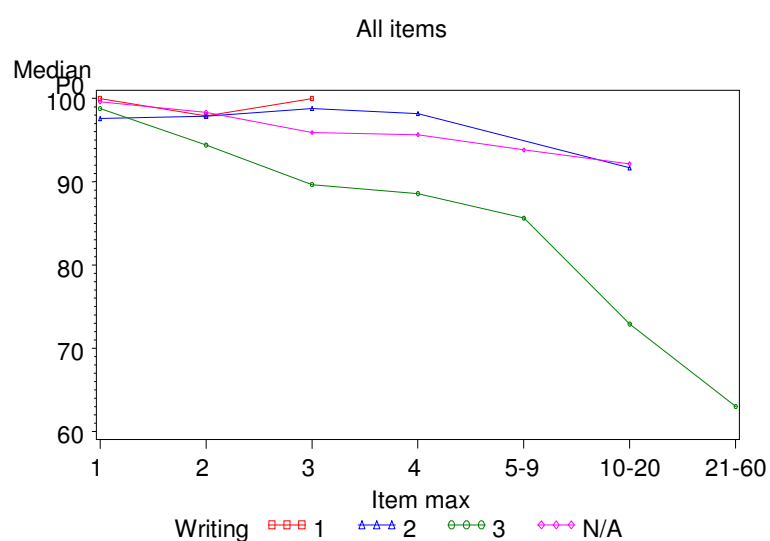


Figure 9: Median P_0 values (all items) for different amounts of writing expected in the response.

In Figure 9 the comparisons based on meaningful numbers of items across the mark range mainly come from items coded '3' or 'N/A' for Writing in the range 2-9 marks. The graph shows that there was much higher agreement (about 6 percentage points) for the 'N/A' items than for items coded '3'. The former were items requiring diagrams, sketches, formulas, equations, arrows, circles, ticks etc. The latter were items requiring two or more sentences. In the logistic regression (Model 8), terms for maximum mark and item type could not be included, because there was too little overlap between the categories for writing, and the different maximum marks. (This is not surprising – questions worth more marks nearly always involve more writing). The effect of each category was compared with 'N/A' (for diagrams, sketches, formulas, equations, arrows etc.). Category '1' (one-word or simple numerical answers) was associated with significantly higher agreement, and category '3' (two or more sentences) was associated with significantly lower agreement. The parameter estimate for category '2' (few words, single sentence) was not significantly different from that for 'N/A'.

Points v levels

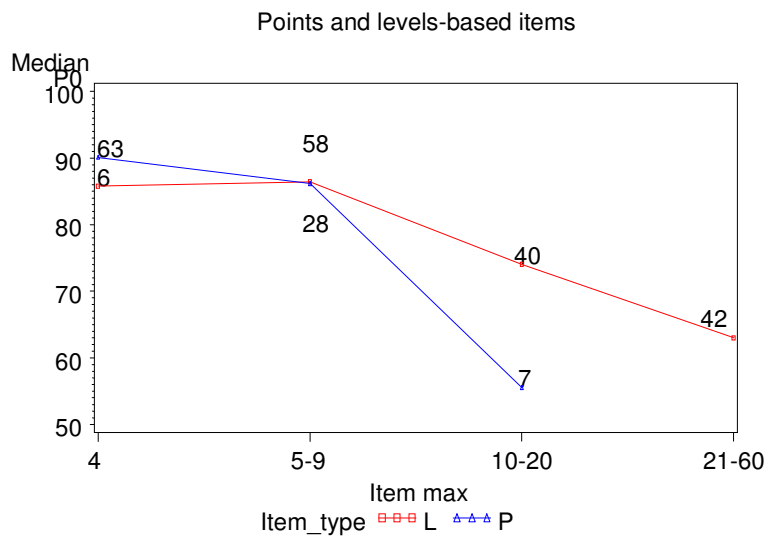


Figure 10: Median P_0 values for points and levels-based items against maximum mark category.

It is interesting to compare the P_0 values for points-based and levels-based items in the mark ranges where they overlap. Figure 10 shows that the median P_0 value was slightly higher for points-based items worth 4 marks, but that the median values were the same for items worth 5-9 marks, and the levels-based items had higher P_0 values for items worth 10 or more marks. In a logistic regression model just containing predictors of maximum mark and item type (Model 9), the main effects and the interactions were statistically significant. The fact that the parameter estimate for the levels-based items was negative suggests that these items were associated with lower agreement than the points-based items, but the fact that the interactions with maximum mark were significant makes it difficult to be conclusive. For example, the positive interaction between 'L' (levels) and max mark suggests that above 6 marks the levels-based items have overcome the disadvantage from the main effect. This shows that it is not necessarily the case that a more 'subjective' mark scheme will lead to less accurate marking. This finding should be treated with some caution however, because the high-mark levels-based items were strongly clustered in particular units (subjects).

Other coded features

The features of Select/Generate and Apparent Marking Strategy Complexity produced results in the expected direction, but are not shown here because for both features there was a large imbalance in the frequencies of the coding categories, with the vast majority of items being coded as requiring the candidate to generate a response, and to require a more complex marking strategy.

Discussion

The qualitative features we coded were all shown to be associated with marker agreement to a greater or lesser extent. Are there any implications for question or mark scheme design? This question cannot be answered without considering validity. As Newton (1996) and many others have pointed out, changing the format of questions or mark schemes to increase the reliability of marking may change what is being assessed. In altering a mark scheme to improve the level of marker agreement it would be very easy to reduce the validity.

A (grossly unrealistic) example would be to decide only to accept one answer in a situation where several valid answers are possible – clearly this would greatly reduce the validity of the question even if it did improve marker agreement. Or imagine a 2-mark question that asked candidates to name two types of rock. The mark scheme might say ‘Any two from: igneous, sedimentary, metamorphic’. This question has a points/marks ratio greater than one, which we have shown is associated with lower levels of marker agreement. The question could be changed to ask candidates to name two types of rock other than igneous. The mark scheme would then be constrained to ‘sedimentary’ and ‘metamorphic’. Alternatively, the question could ask for three kinds of rock, changing the mark allocation to 3 and awarding one mark for each type of rock. Either of these would bring the points/marks ratio to one, which would be expected to increase marker agreement (although other things being equal questions worth more marks have lower marker agreement). However, the first might be objected to on the grounds that it is ‘unfair’ on pupils who only know two out of three rocks, one of them being igneous. The second might in some contexts give too much weight to the question.

To make predictions about marker agreement at this very fine level requires understanding of what *causes* variation in marker agreement, rather than what is merely associated with it, which is likely to require further experimental work systematically manipulating different features of questions and mark schemes. The following paragraphs contains some speculative suggestions of how marker agreement on objective and points-based items might be considered in terms of the probability of an ‘execution error’ in a cognitive processing task.

If the decision to award each mark reflects a single process with a constant probability of error, then the proportion of exact agreement on an n -mark question should be equal to the proportion of exact agreement on a 1-mark question raised to the power n . Table 4 shows these expected proportions for objective and points-based items separately.

Table 4. Observed and expected proportions of agreement for objective and points-based items.

		Item maximum mark					
		1	2	3	4	5	6
Objective	# items	218	61	27	18		
	observed	0.994	0.983	0.969	0.970		
	expected	0.994	0.988	0.982	0.976		
Points	# items	110	206	112	63	21	17
	observed	0.967	0.944	0.920	0.897	0.857	0.850
	expected	0.967	0.934	0.903	0.873	0.843	0.815

The agreement between the observed and expected proportions is quite close, especially for the objective items. This suggests that considering the award of each mark as an independent process with a constant probability of incorrect execution is a reasonable ‘baseline’ model. The fact that the agreement for points-based items is slightly higher for an n -mark task than for n 1-mark tasks is interesting. It seems plausible to assume that there is less of a shift of ‘task set’ (e.g. Allport *et al.*, 1994; Rogers & Monsell, 1995) when carrying

out multiple tasks in the same semantic context than when carrying them out across contexts, and this could be related to the probability of an execution error occurring.

The difference between 'objective' and 'points-based' items as defined here is based on constraint. This is likely to affect the marking strategy used. The simpler strategies of 'matching' and 'scanning for simple items' are more likely in general to be applicable to items with highly constrained mark schemes. The greater automaticity of these strategies presumably implies that they are more likely to be executed without error, and hence that the agreement will be higher, even once number of marks has been taken into account.

A points/marks ratio greater than one can also be seen as increasing the complexity of a given processing task. In the 'types of rock' example above, if it is assumed that a 'matching' strategy is appropriate, then for the original question ('name two types of rock') the first response from the candidate has to be matched against 'igneous', 'sedimentary' and 'metamorphic', and the second response has to be matched against either all three (if the first response was not one of the three correct types) or whichever two remained (if the first response was one of the three types). For the modified question ('name two types of rock other than igneous') the number of correct answers to match the candidate response against has been reduced. If there is a finite probability of an execution error at each matching step then this would lead to higher marker agreement in the second case.

Qualifications, restrictions and variants in the mark scheme (here including wrong answers specifically mentioned) could help when applying the more complex (system-2) marking strategies such as 'evaluating' or 'scrutinising' by increasing the information available to the AE and ensuring that their decision matches the (assumed correct) decision of the TL. However, it might be that this extra information interferes with the more simple (system-1) strategies of 'matching' and 'scanning'. One possibility is that the presence of variant responses forces the marker to use a different cognitive strategy (e.g. 'matching' as opposed to 'scanning') and that this switch carries with it an increased probability of error (for example if the marker had got into an automatic routine of 'scanning' for the correct response and then did not notice when a response was different from the one being scanned for yet nevertheless matched a QRV in the mark scheme). This would fit with the finding that QRVs were associated with higher agreement on points-based items, but lower agreement on objective items.

It is more difficult to relate marker agreement on levels-based questions to the probability of an execution error in a cognitive strategy because it is more difficult to argue that the TL mark (or any one person's mark) is correct. Overall patterns of marker variation are better handled statistically within a many-facet IRT model, or a generalisability theory model, which separate out leniency/severity and erraticism (Bramley, 2007). These models do not say anything, however, about the processes within an individual which lead to the award of a mark. Presumably some kind of matching process is going on in some instances (e.g. those with 'best fit' judgments), but it is very doubtful that this is the same kind of system-1 'matching' referred to above. Also, it is plausible that the TL monitoring role is somewhat different when second-marking essays with a levels-based mark scheme, as opposed to shorter points-based items. In the latter, it might be clear to them that their AE has applied the mark scheme incorrectly, whereas in the former they might be prepared to tolerate differences within a certain range and not award a different mark from the AE unless they seriously disagreed.

We can speculate that the lower marker agreement for items requiring a longer written response might be due to the greater interpretation required by the marker to form a representation of the response which can be compared to the mark scheme. In other words, the marker is likely to encounter more ways of expressing the same concepts and thought processes in writing than in (for example) formulas and equations.

Caveats and limitations

When carrying out the qualitative coding of the question papers and mark schemes we were working from the final version of the question papers, and the latest version of the mark scheme that we were able to obtain. There was some inconsistency across different units in what mark scheme was available. In some cases, it is likely that changes made to the mark scheme at the standardisation meeting⁵ would not have appeared on the versions we coded. This is likely to have affected some of the coding categories more than others – for example it is plausible that more items would have been coded positively for QRV and Wrong if we had had access to the final definitive mark scheme used by the markers.

It is possible that not all scripts would have been available for sampling – for example if they were involved in result enquiries or appeals processes. It is likely that such scripts would have formed a very small proportion of the total, but we might expect such scripts to have shown more evidence of discrepancies between AE and TL. If this is true, then the statistics reported here would overestimate very slightly the true levels of agreement.

This live setting gave the advantage of no possible artefacts (e.g. time lags, the need for extra or special training, the use of photocopied scripts) which might be introduced in a specialised ‘research’ setting. On the other hand, it removed the opportunity for experimental control of the different features of question papers and mark schemes that were coded. We relied on the fact that the sample of units was large and representative of written papers in general qualifications.

Conclusion

This research has shown that some general features of examination question papers and mark schemes, which can be relatively objectively coded across a wide range of subjects, are related to the level of agreement between two markers (or marking accuracy, if one of the marks can be taken as the ‘correct’ mark). In terms of statistical prediction of marker accuracy for objective and points-based items worth up to nine marks, information about item type provides extra information above maximum mark, and information about the points/marks ratio provides extra information above maximum mark and item type. This could be useful in deciding how to allocate resources where there is the option to assign different types of marker to different types of question. In terms of understanding the underlying causes of variation in marker accuracy, these findings fit into a framework that looks to relate question features to cognitive task complexity and to cognitive marking strategies.

References

- Allport, D.A., Styles, E.A. and Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In: C. Umiltà and M. Moscovitch, Editors, *Attention and performance Vol. XV*, A Bradford Book, Cambridge, pp. 421–452.
- Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication*, 4, 22-28.
- Crisp, V. (2008) Improving students’ capacity to show their knowledge, understanding and skills in exams by using combined question and answer papers. *Research Papers in Education*, 23(1), 69-84.
- Jankowicz, D. (2004). *The easy guide to repertory grids*. Chichester: John Wiley & Sons.

⁵The point in the process when final clarifications and amendments are made to the mark scheme, in the light of the PE’s marking of a sample of actual candidate responses.

- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and Biases* (pp. 49–81). Cambridge: Cambridge University Press
- Kramer, A., Coles, M. & Logan, G. (1996). *Converging operations in the study of visual selective attention*. Washington DC: American Psychological Association.
- Kelly, G.A. (1955). *The psychology of personal constructs*. New York: Norton.
- 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.
- Newton, P. (1996). The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal*, 22, 405-420.
- Massey, A.J. & Raikes, N. (2006). *Item level examiner agreement*. Paper presented at the 2006 Annual Conference of the British Educational Research Association, 6-9 September 2006, University of Warwick, UK.
- QCA (2007). *GCSE, GCE, VCE, GNVQ and AEA Code of Practice*. London: QCA.
- Rogers, R.D. & Monsell, S. (1995) Costs of a predictable switch between simple cognitive tasks, *Journal of Experimental Psychology: General*, 124, 207–231.
- Suto, W.M.I. & Greatorex, J. (2008). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*, 34(2), 213-233.
- Suto, W.M.I. and Nadas, R. (2007). 'The 'Marking Expertise' projects: Empirical investigations of some popular assumptions' *Research Matters: A Cambridge Assessment Publication*, 4, 2-5.
- Suto, W.M.I. and Nadas, R. (a, in press). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education*.
- Suto, W.M.I. and Nadas, R. (b, in press). Why are some GCSE examination questions harder to mark accurately than others? Using Kelly's Repertory Grid technique to identify relevant question features. *Research Papers in Education*.

ⁱ The Cambridge Assessment Group is a department of the University of Cambridge and a not-for-profit organisation. It comprises three exam boards. Oxford, Cambridge and RSA (OCR) provides qualifications in the UK. University of Cambridge International Examinations (CIE) provides international 14-19 qualifications. University of Cambridge ESOL Examinations (English for Speakers of Other Languages) offers exams for learners of English.

Appendix A – Statistical modelling

Table A1: Variables included in the logistic regression models.

Model	Items included*	# items	Independent variables
1	Objective and points-based items worth up to 9 marks.	883	Item_max
2	Objective and points-based items worth up to 9 marks.	883	Item_max, item_type
3	Objective and points-based items worth up to 9 marks that had not been coded 'N/A' for PM_ratio.	761	Item_max, item_type, PM_ratio
4	Objective and points-based items worth up to 9 marks that had not been coded 'N/A' for QRV.	882	Item_max, item_type, QRV
5	Objective and points-based items worth up to 9 marks that had not been coded 'N/A' for Wrong.	882	Item_max, item_type, Wrong
6	All items worth up to 9 marks	920	Item_max, ans_space
7	All items worth up to 9 marks	920	Writing
8	Points-based and levels-based items worth between 4 and 20 marks.	202	Item_max, item_type
9	Objective and points-based items worth up to 9 marks that had not been coded 'N/A' for PM_ratio, QRV or Wrong	761	Item_max, item_type, PM_ratio, QRV, Wrong

*All models only included items which had 10 or more data points – i.e. for which we had 10 or more AE/TL marks.

Table A2: Parameter estimates from logistic regression models 1 to 6.

	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Intercept	4.8299	0.0588	4.9649	0.0810	5.0414	0.1135	4.8086	0.0916	4.6058	0.1186	4.7435	0.0650
Item_max	-0.9066	0.0360	-0.8883	0.0593	-0.9860	0.0927	-0.8335	0.0637	-0.7544	0.0766	-0.6889	0.0399
Item_max ²	0.0695	0.00473	0.0805	0.00934	0.1014	0.0164	0.0761	0.0096	0.0692	0.0111	0.0505	0.00496
Item_type O			1.1160	0.0810	0.8546	0.1289	1.0746	0.0815	0.8835	0.0869	0.6700	0.0296
PM_ratio M					-0.4965	0.0962					-0.2401	0.0216
QRV N							0.3160	0.0756				
Wrong N									0.3586	0.1080		
Item_max*Item_type O			-0.3110	0.0593	-0.1941*	0.0996	-0.3229	0.0589	-0.2263	0.0605		
Item_max*PM_ratio M					0.1905	0.0593						
Item_max*QRV N							-0.1416	0.0435				
Item_max*Wrong N									-0.0813*	0.0649		
Item_max ² *Item_type O			0.0439	0.00934	0.0342	0.0170	0.0473	0.00921	0.0334	0.00944		
Item_max ² *PM_ratio M					-0.0279	0.00791						
Item_max ² *QRV N							0.0176	0.00532				
Item_max ² *Wrong N									0.00318*	0.00864		
Item_type*PM_ratio M					-0.0152*	0.0316						
Item_type*QRV N							0.0898	0.0243				
Item_type*Wrong N									0.1646	0.0264		

*parameter estimate is NOT significantly different from zero at the 5% level.

Note that in the above table the parameters for item type, points/marks ratio, QRV and Wrong are estimated for the codings O (objective), M (more than 1:1), and N (not present), with reference categories P (points-based), S (same 1:1), and Y (present) respectively.

Table A3: Parameter estimates from logistic regression models 7 to 9.

	Model 7		Model 8		Model 9	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Intercept	4.2156	0.3303	3.6878	0.0353	3.2772	0.1382
Item_max	-1.0578	0.3377			-0.3186	0.0293
Item_max ²	0.2167	0.1006			0.00977	0.00132
Ans_space 1	3.3747	0.7292				
Ans_space 2	0.1305*	0.3361				
Ans_space 3	-3.3529	0.7398				
Item_max* Ans_space 1	-2.5981	0.9809				
Item_max* Ans_space 2	0.2885*	0.3399				
Item_max* Ans_space 3	1.6688	0.4183				
Item_max ² * Ans_space 1	0.6475	0.3008				
Item_max ² * Ans_space 2	-0.1544*	0.1007				
Item_max ² * Ans_space 3	-0.2888	0.1030				
Writing 1			1.6194	0.0981		
Writing 2			-0.0971	0.0469		
Writing 3			-1.4125	0.0377		
Item_type L					-0.2822	0.1382
Item_max*Item_type					0.0780	0.0293
Item_max ² *Item_type					-0.00379	0.00132

*parameter estimate is NOT significantly different from zero at the 5% level.

Note that in the above table the parameter estimates for Answer space and Writing are with reference to the baseline category of 'N/A'. The parameter for item type L (levels) is with reference to the points-based category (P).

Appendix B – Cognitive marking strategies

The text below is a shorted version of that given in Suto & Greatorex (2008). A similar version to the text below was presented by Suto and Greatorex at the AEA-Europe conference in 2005, available at

<http://www.aea-europe.net/userfiles/DG1%20What%20goes%20through%20a%20Markers%20mind.pdf>

Findings

Five cognitive strategies were identified and interpreted within dual-processing theories of human judgment. Such theories distinguish two qualitatively different but concurrently active systems of cognitive operations: *System 1* or 'intuitive' thought processes, which are quick, associative, effortless and opaque; and *System 2* or 'reflective' thought processes, which are slow, serial, effortful, rule-governed and conscious (Kahneman and Frederick, 2002).

Matching can be used when the answer to a question part is a visually recognisable item or pattern, for example, a letter, word, number, part of a diagram, or short sequence of words or letters. The examiner looks at a short answer line or another pre-determined location and compares the candidate's response with the correct answer (either held in the memory or recollected using the mark scheme), making a simple judgement about whether they match. Matching is an established cognitive process that has already been investigated in many other psychological research contexts. It is likely to entail primarily System 1 thought processing.

Scanning occurs when an examiner scans the whole of the space in the script allocated to a question part, in order to identify whether a particular detail in the mark scheme is present or absent. This detail might be a letter, word, part of a diagram or similar. The cognitive process of scanning is well-established in psychological studies of attention (Kramer *et al.*, 1996). When the scanned-for detail is simple (for example, a single number or letter), pattern recognition occurs and System 1 thought processing appears to be engaged. When the scanned-for detail requires more meaningful or semantic processing, for example, a stage of mathematical working, an additional cognitive strategy entailing System 2 thought processing might need to be used.

When *evaluating*, an examiner attends to either all or part of the space dedicated to an item. He or she processes the information semantically, considering the candidate's answer for structure, clarity, factual accuracy and logic or other characteristics given in the mark scheme. Sometimes a single judgement about the mark value for a particular answer is made at the end of evaluating a response. At other times one or more interim judgements of the appropriate mark value for the candidate's answer are made during the evaluation process. Evaluating is thought to rely primarily on System 2 thought processing.

Scrutinising follows on from, or is used in conjunction with, one of the other cognitive strategies, and is used only when a candidate's answer is unexpected or incorrect. The examiner tries to identify where the problem lies and whether a valid alternative to the mark scheme solution has been given. To do this, he or she evaluates multiple aspects of the candidate's response with the overarching aim of reconstructing the candidate's line of reasoning or working out what the candidate was trying to do. The examiner might have to deal with a lot of uncertainty and re-read the candidate's response several times. As with evaluating, the examiner's thought processes are most likely to be those of System 2.

The *no response* strategy is used when a candidate has written nothing in the answer space allocated to the question part. The examiner looks at the space once or more to confirm this; he or she can award 0 marks for that item.

Appendix C – Examples of how some of the coding categories were applied

C1: Points to marks ratio

The question below was coded as **M** (More) because there were more distinct acceptable points than marks available.

Question:

1 (a) Study Fig. 1, a scatter graph which shows the birth and death rates of seven countries in 2004.

(iv) Suggest reasons why Botswana has a higher death rate than the USA. **[3]**

Mark Scheme:

(iv) Ideally answer should be comparative, however be prepared to link points from separate accounts.

Ideas such as:

better quality health care in USA;

more likely to be preventative measures in USA/vaccination;

better diet/food supply in USA/less likelihood of starvation;

better sanitation in USA;

cleaner water supply in USA;

healthier lifestyle in USA;

AIDS is more of a problem in Botswana;

Education re. health care etc

3 @ 1 mark or development

(3)

The following question was coded as **S** (same) because the number of substantive valid points (ignoring slight variations in wording) was equal to the number of marks available. It also contains an example of a wrong answer specifically mentioned.

Question:

Q3 (c) Explain in detail how carbon monoxide, produced in this reaction, is poisonous. **[2]**

Mark Scheme:

(c) (CO is poisonous...)

due to complexing / ligand exchange with (Fe of) haemoglobin (1)

(NOT redox involving Fe^{2+} / Fe^{3+})

stopping O₂ being transported around body/in blood/to tissues/from lungs (1) **[2]**

C2: Qualifications, Restrictions and Variants (QRV)

The following two questions were coded **Y** (Yes) for the presence of QRVs. The first one also contains an example of an explicit wrong answer (A stands for 'accept' and R stands for 'reject'), so would also have been coded Y for Wrong. The second example allows 'error carried forward' (ecf).

Question:

2 Fig. 2.1 shows a transverse section of a root nodule of a legume. Fig. 2.2 is a drawing of a cell from the centre of the nodule made from an electron micrograph.

- (a) Name three structures that are present in cells in the cortex of the root that are not present in bacterial cells. (3)

Mark scheme:

2 (a) nucleus/nuclear membrane/nuclear envelope/nucleolus;
ER/SER/RER;
Golgi (body/apparatus) / lysosomes;
larger ribosomes/80S ribosomes;
linear DNA/chromosomes/protein + DNA (in chromosomes);
mitochondrion/mitochondria;
cell wall made of cellulose; **R** cell wall unqualified
microtubules; **A** spindle fibres / centriole
large vacuole / tonoplast;
plasmodesmata;
[max 3]

Question:

- 4(b) (iv) Calculate the total energy transformed by the three lamps in kilowatt hours when operated for 12 hours

Mark scheme:

4(b)(iv)
energy = $0.018 \times 12 \times 3$ C1
energy = $0.648 \approx 0.65$ (kW h) (Possible ecf) A1

(0.22 (kW h) scores 1/2)
(648 (kW h) scores 1/2)
(2.3×10^6 (J) scores 1/2)

C3: Wrong (a wrong answer specified in the mark scheme)

The following question was coded **Y** (Yes) for the 'Wrong' category:

Question:

2 Répondez:
À quelle occasion a-t-elle envoyé les fleurs? [1]

Mark Scheme:

Q2
pour son anniversaire de mariage [1]
Reject: *anniversaire* t.c.
Reject: *anni versaire* - two words

Appendix D – Frequency tables

The tables show the number of items contributing to each median plotted in Figures 2 to 10.

Figure 2: Frequencies for objective (O) and points-based (P) items

Item type	Item maximum mark					Total
	1	2	3	4	5-9	
O	218	61	27	18	10	334
P	110	206	112	63	58	549
Total	328	267	139	81	68	883

Figure 3: Frequencies for objective and points-based items with the same (S) and more (M) points than marks.

P:M ratio	Item maximum mark					Total
	1	2	3	4	5-9	
M	96	142	70	51	45	404
S	217	86	31	14	9	357
Total	313	228	101	65	54	761

Figure 4: Frequencies for points-based items with (Y) and without (N) any QRVs.

QRV	Item maximum mark					Total
	1	2	3	4	5-9	
N	96	171	81	48	37	433
Y	14	35	31	15	20	115
Total	110	206	112	63	57	548

Figure 5: Frequencies for objective items with (Y) and without (N) any QRVs.

QRV	Item maximum mark					Total
	1	2	3	4	5-9	
N	190	44	20	17	7	278
Y	28	17	7	1	3	56
Total	218	61	27	18	10	334

Figure 6: Frequencies for points-based items with (Y) and without (N) any wrong answers specified in the mark scheme.

Wrong	Item maximum mark					Total
	1	2	3	4	5-9	
N	100	168	91	59	49	467
Y	10	38	21	4	8	81
Total	110	206	112	63	57	548

Figure 7: Frequencies for objective items with (Y) and without (N) any wrong answers specified in the mark scheme.

Wrong	Item maximum mark					Total
	1	2	3	4	5-9	
N	208	53	17	13	7	298
Y	10	8	10	5	3	36
Total	218	61	27	18	10	334

Figure 8: Frequencies for items with different amounts of answer space.

Ans. space	Item maximum mark							Total
	1	2	3	4	5-9	10-20	21-60	
1	178	10	1	0	0	0	0	189
2	130	216	100	58	29	1	0	534
3	0	0	2	0	14	6	0	22
N/A	21	41	36	29	55	43	42	267
Total	329	267	139	87	98	50	42	1012

Figure 9: Frequencies for items requiring different amounts of writing.

Writing	Item maximum mark							Total
	1	2	3	4	5-9	10-20	21-60	
1	86	1	1	0	0	0	0	88
2	140	27	7	5	0	1	0	180
3	1	151	72	53	76	46	42	441
N/A	102	88	59	29	22	3	0	303
Total	329	267	139	87	98	50	42	1012

Figure 10: Frequencies for points and levels-based items.

Item type	Item maximum mark				Total
	4	5-9	10-20	21-60	
P	63	58	7	0	128
L	6	28	40	42	116
Total	69	86	47	42	244