This is a single article from *Research Matters: A Cambridge Assessment publication*. http://www.cambridgeassessment.org.uk/research-matters/
© UCLES 2009

Mark scheme features associated with different levels of marker agreement

Tom Bramley Research Division

This is a shortened version of a paper presented at BERA in 2008. It does not include the statistical modelling of the results. See Bramley (2008) for full details.

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; Massey and Raikes, 2006). 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 (Suto and Nádas, 2008a, b, *in press*) at Cambridge Assessment 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 and Nádas (2007).

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 Nádas, whose work involved detailed subject-specific analysis in only two subjects (GCSE Maths and Physics). 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 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.

The data came from 38 public examinations in mainstream subjects taken at age 16–18 from OCR (GCSE, AS and A-level in June 2006), and CIE (IGCSE, O-level and A-level in November 2006). In contrast to the research cited above, these data were collected from the process of marker monitoring in the live examinations, as opposed to a research exercise taking place later.

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 data used in this study came from sampling from these re-marked scripts across each team (panel) of examiners. Once the scripts had been obtained, the marks awarded by AE and TL at item level were keyed into a database.

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[1] × an average of 100 candidates per unit × an average of 30 items per unit = 114000).

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.

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 the Appendix for some examples of how the coding framework was applied.

Maximum mark [item_max][2]

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 processing tasks the marker needs to accomplish in marking it. We would probably expect less agreement between markers on questions worth more marks.

Item type [item_type]

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

[1] Here a 'unit' means a single examination paper – usually just one component of several in the complete assessment.

[2] The abbreviation for each category given in square brackets is the variable name which appears in some of the tables and graphs elsewhere in the report.

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 and 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 and Nádas (b, *in press*) found that 'Mark scheme flexibility' and 'Single letter answer' were related to marking accuracy in GCSE Physics (in the expected direction).

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 and Nádas (b, *in press*) found it 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 1, this was not always an

Table 1: Coding framework used to code different features of the question papers and mark schemes[3]

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).
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.

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.

[3] More features than this were coded, but only those features referred to later are listed. See Bramley (2008) for full details.

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 and Nádas (b, *in press*) found that a similar feature of 'alternative answers' was 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. The QRVs might be a help for the larger questions, but a hindrance for the shorter questions. One particular example is where the mark scheme allows 'error carried forward' (ecf)[4]. Suto and Nádas (b, *in press*) 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 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 (Bramley, 2007). 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.

The P_0 statistic was calculated for each item in each unit for which there were more than 10 data points. 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 2 below.

Table 2: 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

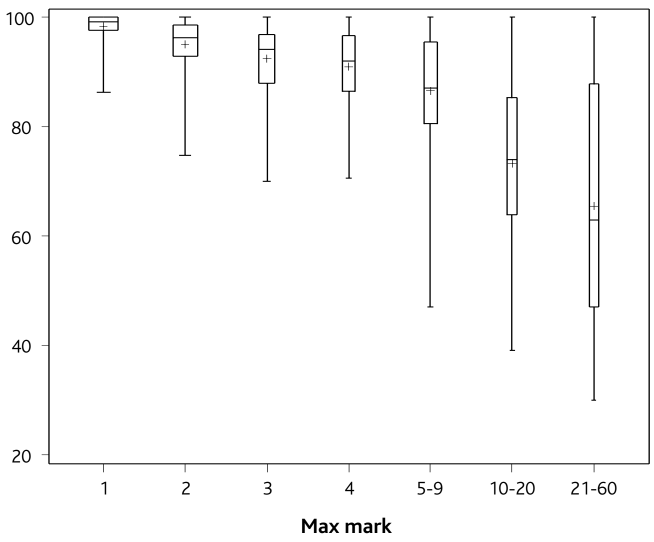

Figure 1: Distribution of P_0 values by item maximum mark. (Width of box is proportional to number of items in each mark category)

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 P0 value in the range ≈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. Many of the coded features were only applicable to objective and points-based items. These items tended to be worth 9 marks or fewer.

Item type

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. 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 and Raikes (2006).

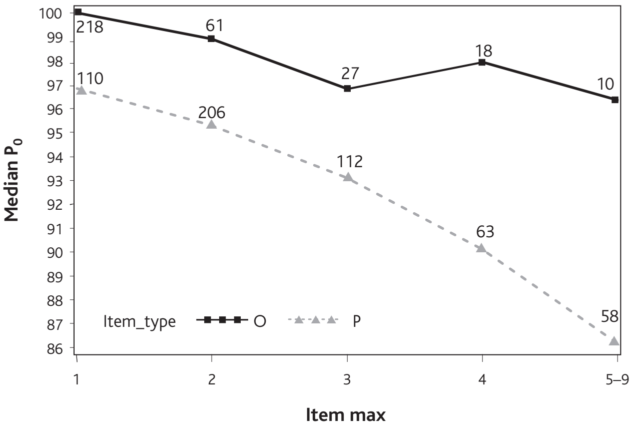

Figure 2: Median P_0 values for objective (O) and points-based (P) items

[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.

Points-to-marks ratio (PM_ratio)

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.

Objective and points-based items

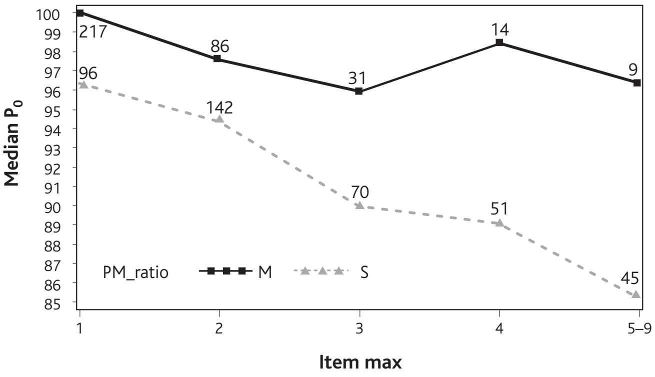

Figure 3: Median P_0 values for objective and points-based items with the same (S) and more (M) points than marks

Points-based items only

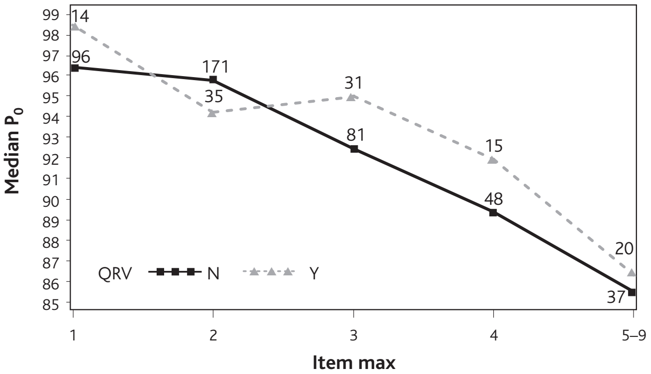

Figure 4: Median P_0 values for points-based items with (Y) and without (N) any QRVs

Objective items only

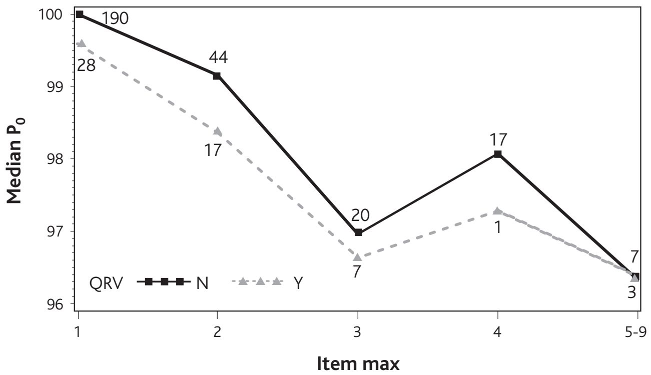

Figure 5: Median P_0 values for objective items with (Y) and without (N) any QRVs

Qualifications, restrictions and variants (QRV)

Figures 4 and 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 very slightly (note the change of scale on the y-axis), as shown in Figure 5. See the discussion for a possible explanation of this result.

Wrong answer specified (wrong)

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.

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.

Points-based items only

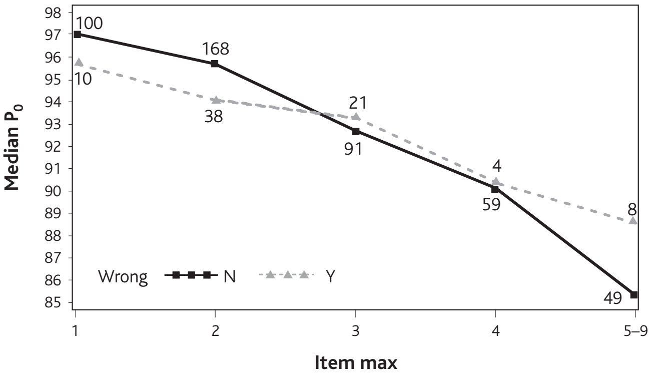

Figure 6: Median P_0 values for points-based items with (Y) and without (N) any wrong answers specified in the mark scheme

Objective items only

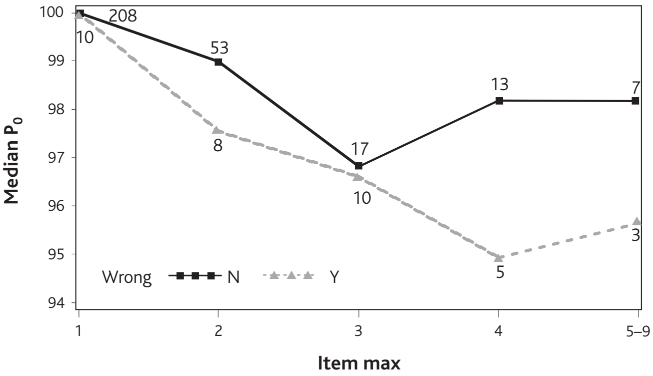

Figure 7: Median P_0 values for objective items with (Y) and without (N) any wrong answers specified in the mark scheme

Answer space (ans_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 – that is, 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 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.

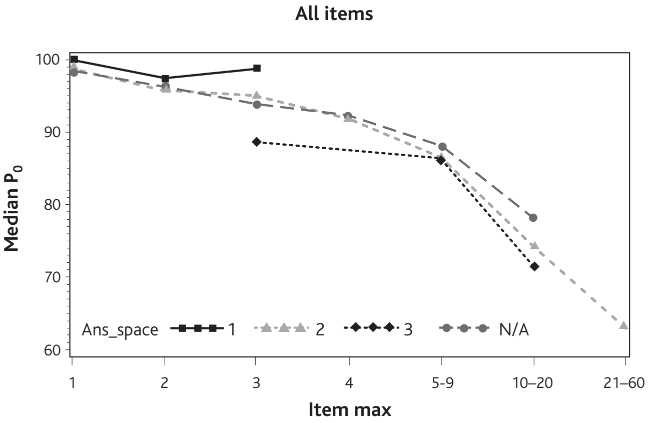

Figure 8: Median P_0 values (all items) for different amounts of answer space

Writing (writing)

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.

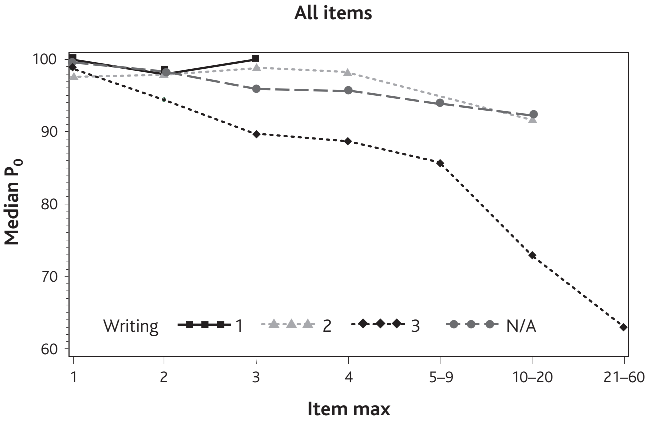

Figure 9: Median P_0 values (all items) for different amounts of writing expected in the response

Points v levels

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. 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).

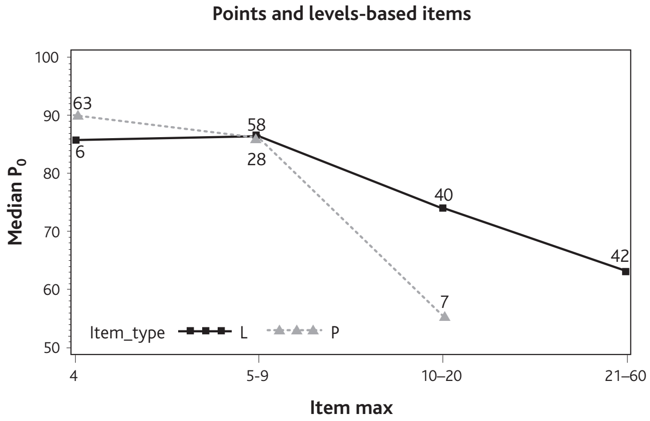

Figure 10: Median P_0 values for points and levels-based items against maximum mark category

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 contain 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 and 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' (Suto and Greatorex, 2008) 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 the 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, we might tentatively assume that: i) 'matching' is an appropriate marking strategy; and ii) that it is a serial process rather than a parallel one. 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 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 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 most common correct response and then did not notice when a correct response was different from the one being scanned for, yet nevertheless matched a QRV in the mark scheme, they would wrongly mark it as incorrect. 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' judgements), but this is not the same kind of '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.

Two caveats in interpreting these results should be mentioned: i) 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 meeting[5] 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; and ii) the 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.

[5] The 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.

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.

In 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). 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. & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In: C. Umiltà and M. Moscovitch (Eds.), *Attention and performance Vol. XV*. Cambridge: Bradford, 421–452.

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication*, **4**, 22–28.

Bramley, T. (2008). *Mark scheme features associated with different levels of marker agreement*. Paper presented at the annual conference of the British Educational Research Association (BERA), Heriot-Watt University, Edinburgh. Available at http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Conference_Papers . Accessed 06/03/09.

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.

Massey, A.J. & Raikes, N. (2006). *Item level examiner agreement*. Paper presented at the annual conference of the British Educational Research Association (BERA), University of Warwick, UK. Available at http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Conference_Papers Accessed 09/03/09.

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.

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. & Nádas, 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. & Nádas, R. (2008a). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education*, **23**, 4, 477–497.

Suto, W.M.I. & Nádas, 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*.

APPENDIX – EXAMPLES OF HOW SOME OF THE CODING CATEGORIES WERE APPLIED

1: 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]**

2: 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 × 12 × 3 C1
energy = 0.648 = 0.65 (kW h) (Possible ecf) A1

(0.22 (kW h) scores ½)
(648 (kW h) scores ½)
(2.3 × 106 (J) scores ½)

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

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

Question:

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

Mark Scheme:

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

ASSESSMENT JUDGEMENTS

Thinking about making the right mark: Using cognitive strategy research to explore examiner training

Dr Irenka Suto, Dr Jackie Greatorex, and Rita Nádas Research Division

This article is based on a presentation, "Exploring how the cognitive strategies used to mark examination questions relate to the efficacy of examiner training", given by Jackie Greatorex, Rita Nádas, Irenka Suto and John F. Bell at the European Educational Research conference, September 2007, Ghent, Belgium.

Introduction

In England, school-leavers' achievements are assessed through a system of public examinations, taken primarily at ages 16 and 18 (Broadfoot, 1996). High stakes examinations for General Certificate in Secondary Education (GCSE) and Advanced (A) level qualifications are administered by three independent awarding bodies, and are marked externally by professional examiners rather than within schools (Ofqual, 2008). Since employers and higher education institutions use GCSE and A-level grades in their selection procedures (Lamprianou, 2008), it is imperative to ensure that examination marking is valid and reliable. This is a considerable task, given the wide variety of question structures and response formats entailed (Eckstein and Noah, 1993). Awarding Bodies therefore conduct rigorous checks on their marking processes and organise highly specialised examiner training, for example in the form of 'standardisation' or 'co-ordination' meetings (National Assessment Agency, 2008). In this article, we investigate the benefits of, and some possible variations in, these training procedures.

GCSE and A-level assessments are in a period of transition. In this context and beyond there has been particular interest in new developments such as on-screen marking (Hamilton, Reddel and Spratt, 2001; Whetton and Newton, 2002; Leacock and Chodorow, 2003; Raikes and Harding, 2003; Sturman and Kispal, 2003; Sukkarieh, Pulman and Raikes, 2005; Knoch, Read and von Randow, 2007; Raikes and Massey, 2007) and the employment of examiners with differing levels of teaching and examining experience (Powers, Kubota, Bentley, Farnum, Swartz and Willard, 1998; Royal-Dawson, 2005; Raikes, Greatorex and Shaw, 2004; Meadows and Wheadon, 2007; Suto and Nádas, 2007a). The focus on examiners with potentially varying expertise has arisen in part because the UK has recently faced shortages of experienced examiners (usually experienced schoolteachers) in some subjects. Moreover, on-screen