# Features of a Levels-Based Mark Scheme and their Effect on Marking Reliability

Anne Pinot de Moira

## Abstract

Levels-based mark schemes are commonly used in the marking of extended response items but, between specifications, there is little commonality in their design, formulation and application. This study firstly establishes a list of the variable design characteristics between levels-based mark schemes and then, secondly, analyses marking reliability with reference to these characteristics. It finds that most of the variation in marking reliability is due to the vagaries of individual responses, which a holistic approach to item design might mitigate. It also recommends a number of small adjustments to mark scheme design which might improve marking reliability and increase the transferability of skills between the marking of different items, units and specifications.

*Keywords: Marking reliability; multilevel modelling; levels-based mark schemes*

## Introduction

### Background

With any large scale assessment, there are many variables which can impact upon the reliability of the outcome reported to students. It is beholden upon assessment practitioners to understand the variables inherent within their assessment in order that grades awarded to students accurately reflect ability in the tested domain. Each of the variables, or components of an assessment, will pose different challenges and, therefore, there is no panacea which will resolve all issues of reliability. Question papers, mark schemes, examiners, administrative procedures, teachers, current affairs and students themselves, all have the potential to reduce reliability in the measurement process.

Over the past few years, there has been an increased focus on the role of the mark scheme in reliable assessment. Pollitt & Ahmed (2008) positioned the mark scheme at the centre of the assessment development process; recognising the impact that a poorly teamed question paper and mark scheme could have upon both reliability and validity. They suggested a schema for writing assessments called Outcome Space Control and Assessment (OSCA) (summarised in Figure 1) and provided examples of how this schema might be used (Pollitt, Ahmed, Baird, Tognolini, & Davidson, 2008). In so doing, they started down the path of incremental and continuous improvement. At the same time, small manageable practical recommendations have been made, many of which are documented in the next section.

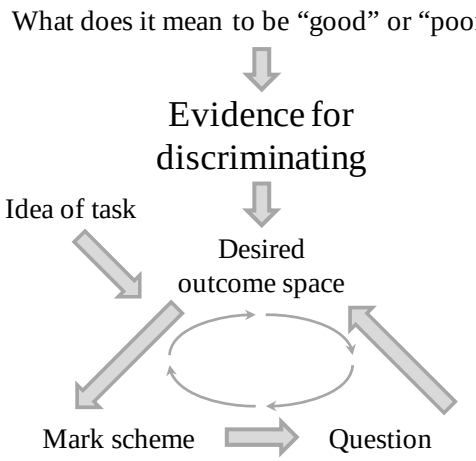


**Figure 1 OSCA, a schema for systematic writing of examination tasks (Pollitt & Ahmed, 2008)**

This paper extends the idea that improvements in the reliability of marking can be made by looking into the detail of the process. Using data from the summer 2011 examination series, it explores the features of levels-based mark schemes with a view to considering the effect these features have on marking reliability. In this context, marking reliability is defined as the level of agreement between two examiners working independently. The study is in two parts. In the first, mark schemes were scrutinised to establish a list of the variable design characteristics and, in the second, marking reliability was analysed statistically with reference to these characteristics.

**Levels-based mark schemes**

Levels-based mark schemes are used predominantly for items with a high mark tariff where there is an extended written response. Such items have scope for multiple valid approaches, rendering point-based marking or the provision of exemplar answers impractical. An examiner is expected to make an initial assessment of a response and, once the response is classified into a single defined level, the examiner is then required to refine this judgement to award a mark (see Figure 2 for an example of a levels-based mark scheme).

| Level | Assessment Objective AO1<br>Knowledge and Understanding |
|---|---|
| Level 3 | **(4–5 marks)**<br>Answers demonstrate a range of citizenship knowledge and an accurate understanding of relevant citizenship concepts and theories.<br>A range of examples is used to relate knowledge and understanding to citizenship issues. |
| Level 2 | **(2–3 marks)**<br>Answers are characterised by a good level of citizenship knowledge and an understanding of relevant citizenship concepts and theories.<br>Examples are used to relate knowledge and understanding to citizenship issues. |
| Level 1 | **(1 mark)**<br>Answers are characterised by limited citizenship knowledge and limited understanding of relevant concepts and theories.<br>Candidates may make a limited attempt to use examples to relate knowledge and understanding to citizenship issues, or no examples may be present. |
| | **(0 marks)**<br>No relevant response. |

**Figure 2 Excerpt from GCE Citizenship Studies Unit 1 generic mark scheme for items 1 and 5 – summer 2011**

While there may be a common understanding of the philosophy behind levels-based mark schemes, there is little commonality in their design and formulation. There are considerable differences in look and feel and these differences must present varying cognitive challenges.

## Features of a levels-based mark scheme

From the many levels-based mark schemes used to mark items on summer 2011 A-level question papers, a sample of over 300 were scrutinised to establish a list of the variable design features. These mark schemes were selected on the basis that the features could be quantified for future modelling and, whilst not an exhaustive list, they represented a high proportion of all AQA large entry A-levels with long form answer questions. Areas of difference are listed below, along with a brief description of the implication of these differences as understood from the current research literature.

1. The number of levels in the mark scheme.

   Clearly, the number of levels is inextricably linked to the maximum marks for an item and therefore to the intended weight of that item (and area of specification content) within the assessment. However, there is an extensive literature, succinctly summarised in Peterson (2000, p. 63), which discusses the optimal number of categories or levels for a rating scale. Despite much discussion of seven as a magic number, it is clear that the task asked of the rater, in this case an examiner, and the use to which the rating is put have an important role in determining the optimal number of levels (Miller, 1956).

2. The number of marks within a level.

   As with the number of levels, decisions regarding the number of marks within a level are linked to the limits of cognitive discrimination and to the desired content weight within the specification.

3. The distribution of marks between levels.

   Theoretical evidence suggests that the number of marks should be equal across all levels described in the mark scheme for an item (Pinot de Moira, 2012).

4. The evaluation, or otherwise, of quality of written communication in the mark levels.

   Quality of written communication (QWC)[1] is most often assessed and evaluated in open ended response items. In the past, a judgement has been made across an entire script but, with the introduction of item level marking, QWC marks are often assigned on the basis of one item alone. In many subjects, QWC has low correlation with the subject specific construct being measured (see for example, Massey & Dexter, 2002). Effective design of a levels-based mark scheme for items where QWC is integrated into the assessment would, therefore, appear to require its separate evaluation outside the levels.

5. The presentation of levels in a grid-like format to separate the evaluation of assessment objectives.

   For some items, the mark scheme makes a distinction between performances on the different assessment objectives tested within. Implicit is the assumption that the correlation between these performances may be low and it would, therefore, be invalid to use a single

[1] Quality of written communication is also sometimes referred to as spelling, punctuation and grammar (SPaG).

levels-based model. Where this is the case, the mark scheme is often presented as a grid with levels forming the rows and assessment objectives forming the columns. The drawback of such a design is that, as the number of cells in the grid increases, so the mark scheme tends towards a points-based system where the award of every mark is specified in detail. It would contradict evidence which suggests that levels-based mark schemes are better, in terms of marking reliability, for items with a maximum tariff of 10 and above (Bramley, 2008).

6. The inclusion, or otherwise, of a mark of zero in the bottom level.

   In some mark schemes the mark of zero (nothing creditworthy) is included in the lowest level and in others it is identified separately outside the levels.

7. The inclusion, or otherwise, of indicative content within the levels.

   In some levels-based mark schemes the level descriptions are generic, while in some they contain indicative content. While there is no research evidence to suggest which design is preferable, the cognitive load would undoubtedly differ dependent upon the wordiness of the mark scheme.

8. The order of presentation of levels: lowest first or highest first.

   Perhaps surprisingly, there is variation in mark schemes as to whether the highest or lowest level is described first. This may introduce a tendency towards positive or negative reward, which differentially influences examiners, especially as they are entreated to be open-minded and positive when marking scripts; crediting what a candidate knows.

9. The documentation, or lack of documentation, to describe the application of the levels-based mark scheme.

   Very few mark schemes include any instructions to examiners on how to use levels-based mark schemes.

While there are undoubtedly pragmatic reasons for variations in mark scheme design, in the interests of improving marking reliability, a clearer understanding of the impact of the varying features would be desirable. This understanding would also facilitate the assembly of a coherent evidence-based guidance for use in assessment development to eliminate arbitrary, and potentially damaging, variations. Matching mark remark data to a sample of levels-marked items gives the opportunity to consider whether marking reliability is influenced to any measureable extent by the mark scheme.

## Modelling Marking Reliability

### Data & model structure

To improve understanding of mark remark reliability, data from 16 units and 133 items were explored using a series of multilevel models (details are given in Appendix A). The data were taken from long form answers that were double marked in summer 2011. They represented an opportunity sample of responses which were remarked for quality control purposes during the marking period. All the items included in the analysis were on-screen marked and this accounts, in part, for attrition from the 300 items originally scrutinised to identify features of levels-based mark schemes. Further attrition resulted from difficulties with encoding QWC when it was assessed separately from the levels. Therefore, all items selected had either no evaluation of QWC or the QWC was embedded within the levels. Although the units were systematically selected, the sample of responses drawn from within these units was, nonetheless, random.

The data were restricted to responses marked within the designated marking period to mitigate against a preponderance of senior examiners in the sample. Before and after the official marking period senior examiners would be over-represented because they are involved with marking set up and with tying up loose ends.

The dependent variable in the model was formulated in one of two ways:

- a binary variable denoting whether the mark awarded by an examiner was in agreement with the final mark awarded (Model 1); and
- a continuous variable which measured the absolute difference between the examiner mark and the mark finally awarded (Models 2 & 3).

The rules for determining the final mark for double marked responses state that, where there is agreement between two examiners within a predetermined tolerance, the original examiner's mark is awarded. Where there is disagreement, defined as a difference greater than the predetermined tolerance, the response is sent for adjudication. The adjudicator, who is normally a senior examiner, will judge which mark of the two is correct and this mark will be chosen as the final mark. If the adjudicator is not happy with either of the marks, the final mark will be of his or her choosing.

For each model, response was nested within item, which was nested within examiner, which was nested within unit; making a four-level model. Models 1 and 2 used all the mark remark data whereas Model 3 extracted only instances where there was a difference between the examiner mark and the final mark. This was in an attempt to combat problems encountered with the sparse dataset. Details of the three models are included in Appendix B.

The set of independent variables were largely derived from the descriptive differences between levels-based mark schemes itemised in 1-3 and 5-8 above. They were augmented by a number of operationally collected variables. These included:

- the unit and item maximum mark so as to control for variations in the extent to which examiners could differ. Theoretically, a low tariff item is almost bound to be marked more accurately. Figure 3 illustrates the probability of agreement between an examiner and the final mark for the data included in the study and provides empirical evidence to support the theory;
- the final mark awarded for the response and the item facility, because previous work has suggested that difficult items prove more difficult to mark (Sweiry, 2012). In the model, final mark is centred to ease interpretation; allowing presentation of findings relative to the mid-point on the mark scale. This variable is also included in the model as a squared value because anecdotally it has been suggested some examiners resist awarding marks at the extremes of the mark range. It is said, in so doing, they believe they will avoid falling foul of a quality control system which bars marking without retraining, given failure to mark double marked items within a set tolerance. Notwithstanding the anecdotal evidence, generally speaking, there are likely to be fewer responses at the extremes, and this could render the marking of these responses less reliable;
- the optionality of an item within the unit. Examiners, as well as candidates, might exhibit preferences or areas of specialism which could affect their ability to mark particular items reliably;
- the time of day that the response was marked; and
- the percentage of the way through the marking period the response was marked. It has been shown, in the past, that there is a small change in the accuracy of marking over time (Pinot de Moira, Massey, Baird, & Morrissy, 2002)

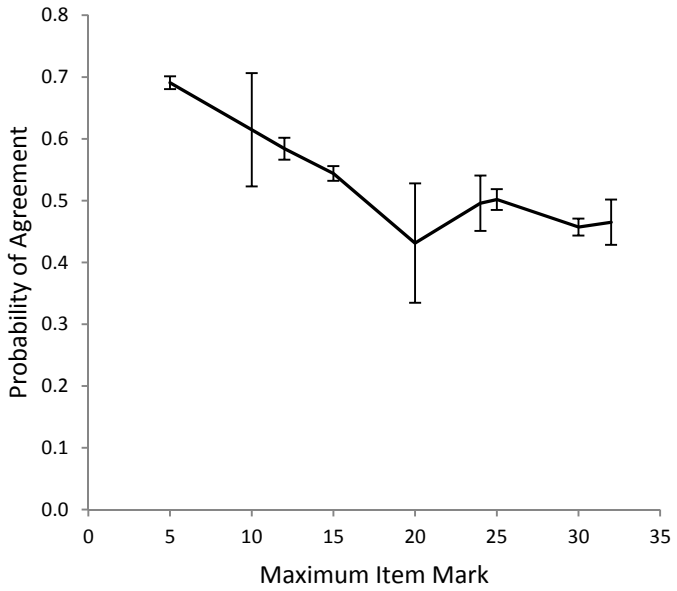


**Figure 3 Probability of agreement between examiners (and associated standard error) using raw data from the 133 items included in the study**

Main effects and interactions were fitted to each model but ultimately the interactions added little extra information and therefore the final structures included only the main effects.

**Findings**

*Explained variation*

The key finding, emerging from all three models, was that features of the mark scheme explained very little variation in the data. Even after all main effects were fitted, there remained considerable unexplained variation at response level and this variation dwarfed that at item, examiner and unit level. The individual responses given by students, therefore, appeared to be the limiting factor for reliable marking.

Whilst this might be regarded as a reason for despair it suggests that, in terms of assessment design, a focus on the interaction between mark scheme and item, rather than the mark scheme alone, might prove more profitable in the quest for improved marking reliability. It is unrealistic to expect, or indeed wish for, the eradication of idiosyncratic responses, but items carefully designed to elicit a consistent approach, where appropriate for valid assessment, will almost certainly improve marking reliability.

*Independent Variables*

Notwithstanding the limitations of the models, there were some consistent patterns across all three. In other words, there were some independent variables which related to the dependent variable in the same way no matter which model was considered. These independent variables were not always significant in a statistical sense and therefore require cautious interpretation. The probabilistic chance of all outcomes coinciding must be considered but, in most cases where there was coincidence, there was some existing research evidence to support the finding.

Both the centred final mark for a response and the squared centred final mark appeared to influence the level of agreement between examiners. With reference to Models 2 and 3, as each variable increased, the difference between the examiner mark and the final mark also increased. Thus, the better the response, the more difficult it was to mark and similarly, the

nearer the extreme of the mark range, the more problems it caused. Whether the difficulties at the extremes were caused by the sparseness of responses, by examiner behaviour modified to game the quality control system or by some other factor is, however, unclear. Using data from Model 3, Figure 4 depicts the relationship between final mark and marking reliability.

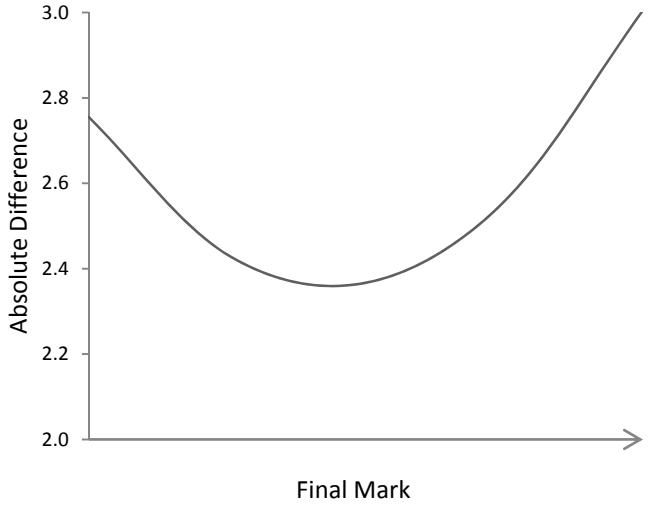


**Figure 4 The relationship between final mark and marking reliability using parameter estimates from Model 3**

The picture with Model 1 was slightly different. As the centred final mark increased, so the probability of agreement between the examiner mark and final mark decreased. However, as the squared centred final mark increased, so the probability of agreement increased. At the extremes of the mark distribution, the potential difference in marks between examiners is at its highest but the probability of agreement is also at its highest. Hence the difference in findings between Model 1, with a binary dependent variable, and Models 2 and 3, with the continuous dependent variable.

Whilst some of the existing literature suggests that it is the difficult items which are problematic to mark (Sweiry, 2012), the models herein appear to show that it is a higher quality of response rather than a greater item difficulty which lowers reliability; an observation also made by Pinot de Moira (2003).

The models all showed that there was an extremely small but statistically significant improvement in marking over the marking period. Examiners were on average a quarter of a mark closer to the final mark by the end of their marking[2]. Evidence, perhaps, that rather than suffering from fatigue as time progressed, the examiner's increasing experience and regular feedback on performance served to hone skills. Another administrative feature that appeared to affect marking reliability was time at which marking was completed. Marking out of school hours (before 8am or after 4pm) tended to be less reliable. According to Model 1, an examiner mark was 1.12 times more likely to coincide with the final mark if it was marked in school hours. For many specifications, the marking period is during school term time and therefore any teachers who are also examiners will be under pressure to mark while still undertaking a full teaching schedule. Examiners marking during school hours are less likely to be current teachers. They are also more likely to be older, retired and more experienced in examining. In other words, the

[2] From Model 2, the parameter estimate for marking period is -0.0024 to four decimal places. Marking period was expressed as a percentage of time elapsed. Therefore, when the marking period was complete, the average mark difference was approximately quarter of a mark lower (-0.0024 x 100) than at the beginning of the marking period.

effect was as likely to be a function of examiner characteristics as the time of day at which the marking was completed.

The remaining three independent variables which suggested a consistent interpretation across all models described features of the mark scheme. The first related to the band descriptions. Marking reliability was higher, although not statistically significantly, when the band descriptions were generic rather than including indicative content specific to the particular item. It might seem counterintuitive that mark schemes with more supporting information result in less reliable marking. However, visual comparison of the generic and specific levels-based mark schemes reveals what may well be the root of the problem. Generic mark schemes are often simple, neat and uncluttered. They are the same in format throughout the unit and, therefore, require less cognitive demand of the user. Levels-based mark schemes which include indicative content in the bands tend to be lengthier and, by definition, differ across the unit.

The second consistent mark scheme feature described, albeit non-significantly, by the models was the effect of the distribution of marks across bands. Previous theoretical research showed that, in order to reduce bias in the distribution of marks across the mark range for an item, the number of marks within each of the levels of a levels-based mark scheme should be the same (Pinot de Moira, 2012). The models showed some support for this finding insofar as they indicated that marking was more reliable when each band was composed of the same number of marks. The extent of this increased reliability appeared to be of the order of half a mark.

Finally, each of the three models suggested that marking was more reliable if the lowest level was described first on the mark scheme. Even allowing for the limitations of the model, the effect size was small and, unlike the other findings, was unsupported by independent literature or simple reasoning.

On many mark schemes, among the general marking guidelines, are instructions to be positive in marking, to award marks which reflect the expected level of performance for the qualification, to use the whole mark range and not to deduct marks for irrelevant or incorrect answers. While doubtless these instructions are not regularly reread, they reflect the philosophy for marking. The interaction of this philosophy with the design of the mark scheme might give clues to the better reliability for mark schemes where the marks are described in ascending order.

On using a mark scheme with the maximum mark at the top of the page and reading downwards, an examiner will be starting from the point of *perfection*. Thus, the examiner is required to deduct rather than to award marks; undermining the established philosophy. Of the 12 units in which the mark scheme detailed the highest level first, two explicitly described the need for positive marking and four required a top down approach to arriving at a final mark. At best, marking philosophy, whether explicitly described or etched in folklore, sometimes seems to be at odds with mark scheme design.

This effect could be seen as analogous to Bramley's (2008) finding which suggested that the addition of qualifications, restrictions and variants (QRVs) to a mark scheme reduced marker agreement. Bramley argued that including QRVs led to examiners switching to more complex cognitive strategies to mark; leading to more errors. Even if QRVs are not included explicitly in levels-based mark schemes, they may still be implicit in examiners' thinking if the top band is presented first.

## Conclusions

### Limitations of the model and potential improvements

Whether the models described herein could be improved is a moot point. The number of units marked at item level within AQA has reached a plateau and there are still many high stakes

subjects with high tariff items which are traditionally marked. To expand the models for greater generalisability would require the introduction of a new swathe of units, with levels-marked items, to the item level marking system. Maybe, on the other hand, improvements could be made if the features of the mark scheme were described subjectively rather than analytically. Bramley (2008), for example, included variables such as the complexity of marking strategy in his model of marking reliability. Furthermore, and at the risk of overanalysing the data, it would be possible to revisit the inclusion of interactions into the models. This might, in particular, shed light on the reasons that good responses prove more difficult to mark.

**Recommendations**

Rather than providing unequivocal evidence to support effective design of levels-based mark schemes, this study serves to highlight the differences in practice that currently exist between specifications. Plainly there is an argument for flexibility in mark scheme design so that the mark scheme suits the subject being assessed. However, there is also an argument for greater commonality to improve reliability and to increase the transferability of skills. There seems to be no rationale for differing marking philosophies and guidelines. Furthermore, it seems logical that we should strive to present mark schemes in a way which minimise the cognitive demand to the examiner. Returning to the areas of difference identified earlier, the following recommendations are made with a view to improving marking reliability:

1 & 2. The number of levels and marks within levels in the mark scheme.

The number of levels in a mark scheme should be determined by the intended weight of the item and by the extent to which the levels can be uniquely described. As with the number of levels, decisions regarding the number of marks within a level should be determined by the limits of cognitive discrimination and to the desired content weight within the specification.

3. The distribution of marks between levels.

As far as possible, the number of marks within each level of a levels-based mark scheme should be equal.

4. The evaluation, or otherwise, of quality of written communication in the mark levels.

Quality of written communication (QWC) should be evaluated separately from the subject-based content and its evaluation should be independent of the levels-based mark scheme.

5 & 7. The presentation of levels in a grid-like format to separate the evaluation of assessment objectives. The inclusion, or otherwise, of indicative content within the levels.

Mark schemes should be designed with cognitive demand in mind. Clear, concise and simple mark schemes are likely to elicit more reliable marking.

9. The documentation, or lack of documentation, to describe the application of the levels-based mark scheme.

Mark schemes, and in particular levels-based mark schemes, should include clear and concise instructions for use. They should promote a consistent philosophy to marking which, in turn, should allow greater transferability of skills between units and specifications.

Above and beyond design of the mark schemes, it seems evident that marking might be improved if time is invested in providing support to examiners who manage their examining workload alongside a teaching schedule. Furthermore, given that marking reliability appears to

have the greatest variation at the individual response level, careful item design might help alleviate marking difficulties. A schema such as that proposed by Pollitt et al (2008) might be used to limit item ambiguity and reduce the multiplicity of responses without compromising the validity of the assessment. At the same time, this focus on the assessment as a whole could be used to consider the effective design of items and mark schemes to discriminate accurately between the higher quality responses.

**Future work**

Although mark schemes only form one part of a successful assessment, there is merit in identifying areas for improvement and future research. Clearly there is a need to understand, and address, why it is harder for examiners to mark good quality responses. It would also be helpful to understand the impact of mark scheme design on cognitive demand. In particular whether, and if so why, the order of presentation of levels in a levels-based scheme affects marking reliability. Linked to this, there is a need to determine the optimal number of levels and marks per level such that an item discriminates effectively between candidates without making unrealistic demands of the examiner.

To a certain extent, the probability of agreement between examiners is determined by the amount to which they *can* disagree (Figure 3). However, hidden amongst the data and the models, are some items which are reliability marked and some which are not. By identifying the former, future work could also involve learning from, and distilling the features of, these examples of good practice.

## Acknowledgements

With thanks to Michelle Meadows and Debra Malpass for their help in the interpretation of the findings from this study.

## References

Bramley, T. (2008). *Mark scheme features associated with different levels of marker agreement*. Presented at the British Educational Research Association (BERA) Annual Conference, Heriot-Watt University, Edinburgh, UK.

Long, J. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage Publications.

Massey, A., & Dexter, T. (2002). *An evaluation of Spelling, Punctuation and Grammar assessments in GCSE*. Presented at the British Educational Research Association (BERA) Annual Conference, Exeter University, Exeter, UK.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, *63*, 81-97.

Peterson, R. A. (2000). *Constructing effective questionnaires*. Thousand Oaks, CA: Sage Publications.

Pinot de Moira, A. (2003). *Examiner Background and the Effect on Marking Reliability* (No. RPA_03_APM_RC_218). Manchester: AQA Centre for Education Research and Policy.

Pinot de Moira, A. (2012). *Levels-based mark schemes and mark bias* (No. CERP_12_APM_RP_015). Manchester: AQA Centre for Education Research and Policy.

Pinot de Moira, A., Massey, C., Baird, J., & Morrissy, M. (2002). Marking consistency over time. *Research in Education*, *67*, 79-87.

Pollitt, A., & Ahmed, A. (2008). Outcome Space Control and Assessment. In *9th Annual Conference of the Association for Educational Assessment – Europe*. Presented at the 9th Annual Conference of the Association for Educational Assessment – Europe, Hissar, Bulgaria.

Pollitt, A., Ahmed, A., Baird, J., Tognolini, J., & Davidson, M. (2008). *Improving the quality of GCSE assessment*. Qualifications and Curriculum Authority. Retrieved from http://www.lifeinbits.org/camexam/htdocs/papers/2008ImprovingQuallityofGCSE.pdf

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis : an introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage Publications.

Sweiry, E. (2012). *Conceptualising and minimising marking demand in selected and constructed response test questions*. Presented at the Association for Educational Assessment (AEA) Europe Annual Conference, Berlin, Germany.

16 May 2013

## Appendix A Sample size for units and items included in the models (summer 2011)

| Unit | Item | Models 1& 2 | Model 3 |
|---|---|---|---|
| CIST1 | 3 | 178 | 68 |
| CIST1 | 4 | 226 | 102 |
| CIST1 | 7 | 130 | 54 |
| CIST1 | 8 | 258 | 102 |
| CIV1D | 5 | 35 | 14 |
| CIV1D | 6 | 34 | 20 |
| CIV1D | 10 | 74 | 28 |
| CIV1D | 11 | 68 | 38 |
| CIV1D | 12 | 64 | 33 |
| CIV1D | 13 | 35 | 21 |
| CRIT2 | B9 | 196 | 106 |
| FREN3 | B10A | 127 | 43 |
| FREN3 | B10C | 127 | 67 |
| FREN3 | B10CL | 127 | 41 |
| FREN3 | B10RV | 127 | 43 |
| FREN3 | B11A | 170 | 49 |
| FREN3 | B11C | 170 | 86 |
| FREN3 | B11CL | 170 | 46 |
| FREN3 | B11RV | 170 | 48 |
| FREN3 | B12A | 414 | 133 |
| FREN3 | B12C | 414 | 216 |
| FREN3 | B12CL | 414 | 135 |
| FREN3 | B12RV | 414 | 137 |
| FREN3 | B13A | 126 | 30 |
| FREN3 | B13C | 126 | 57 |
| FREN3 | B13CL | 126 | 38 |
| FREN3 | B13RV | 126 | 39 |
| FREN3 | B14A | 442 | 145 |
| FREN3 | B14C | 442 | 238 |
| FREN3 | B14CL | 442 | 149 |
| FREN3 | B14RV | 442 | 151 |
| GENB1 | 1 | 495 | 526 |
| GENB1 | 2 | 496 | 414 |
| GENB1 | 3 | 474 | 275 |
| GENB1 | 4 | 471 | 282 |
| GENB1 | 5 | 495 | 358 |
| GEOG1 | A1c | 489 | 976 |
| GEOG1 | A2c | 484 | 325 |
| GEOG1 | A3c | 490 | 662 |
| GEOG1 | A4c | 153 | 75 |
| GEOG1 | B5c | 486 | 904 |
| GEOG1 | B6c | 303 | 136 |
| GEOG1 | B7c | 488 | 332 |
| GEOG1 | B8c | 485 | 482 |
| GERM3 | B11A | 33 | 8 |
| GERM3 | B11C | 33 | 13 |
| GERM3 | B11CL | 33 | 8 |
| GERM3 | B11RV | 33 | 7 |
| GERM3 | B12A | 88 | 25 |
| GERM3 | B12C | 88 | 41 |
| GERM3 | B12CL | 88 | 25 |
| GERM3 | B12RV | 88 | 26 |
| GERM3 | B13A | 81 | 27 |
| GERM3 | B13C | 81 | 38 |
| GERM3 | B13CL | 81 | 18 |
| GERM3 | B13RV | 81 | 20 |
| GERM3 | B14A | 78 | 19 |
| GERM3 | B14C | 78 | 39 |
| GERM3 | B14CL | 78 | 20 |
| GERM3 | B14RV | 78 | 21 |
| GERM3 | B15A | 130 | 41 |
| GERM3 | B15C | 130 | 63 |
| GERM3 | B15CL | 130 | 33 |
| GERM3 | B15RV | 130 | 34 |
| GOVP1 | 3 | 326 | 155 |
| GOVP1 | 6 | 256 | 121 |
| GOVP1 | 9 | 47 | 22 |
| GOVP1 | 12 | 207 | 101 |
| HIS1E | 1 | 43 | 16 |
| HIS1E | 2 | 45 | 23 |
| HIS1E | 3 | 55 | 22 |
| HIS1E | 4 | 55 | 24 |
| HIS1E | 5 | 40 | 18 |
| HIS1E | 6 | 44 | 22 |
| HIS1F | 1 | 99 | 44 |
| HIS1F | 2 | 95 | 43 |
| HIS1F | 3 | 56 | 25 |
| HIS1F | 4 | 61 | 33 |
| HIS1F | 5 | 48 | 24 |
| HIS1F | 6 | 40 | 22 |
| HIS1K | 1 | 46 | 15 |
| HIS1K | 2 | 46 | 23 |
| HIS1K | 3 | 51 | 23 |
| HIS1K | 4 | 51 | 30 |
| HIS1K | 5 | 45 | 23 |
| HIS1K | 6 | 43 | 22 |
| MEST1 | A1 | 497 | 473 |
| MEST1 | A2 | 499 | 404 |
| MEST1 | A3 | 497 | 443 |
| MEST1 | A4 | 499 | 436 |
| MEST1 | B5 | 488 | 380 |
| MEST1 | B6 | 228 | 111 |
| PHED1 | B7 | 470 | 360 |
| PHIL1 | 1 | 481 | 277 |
| PHIL1 | 2 | 472 | 294 |
| PHIL1 | 3 | 129 | 54 |
| PHIL1 | 4 | 118 | 50 |
| PHIL1 | 5 | 193 | 84 |
| PHIL1 | 6 | 172 | 80 |
| PHIL1 | 7 | 175 | 78 |
| PHIL1 | 8 | 178 | 82 |
| PHIL1 | 9 | 104 | 48 |
| PHIL1 | 10 | 90 | 40 |
| PHIL2 | 1 | 328 | 139 |
| PHIL2 | 2 | 317 | 183 |
| PHIL2 | 3 | 177 | 81 |
| PHIL2 | 4 | 162 | 80 |
| PHIL2 | 5 | 221 | 100 |
| PHIL2 | 6 | 202 | 100 |
| PHIL2 | 7 | 356 | 160 |
| PHIL2 | 8 | 349 | 181 |
| PHIL2 | 9 | 313 | 132 |
| PHIL2 | 10 | 277 | 146 |
| SPAN3 | B10A | 84 | 23 |
| SPAN3 | B10C | 84 | 44 |
| SPAN3 | B10CL | 84 | 24 |
| SPAN3 | B10RV | 84 | 26 |
| SPAN3 | B11A | 83 | 29 |
| SPAN3 | B11C | 83 | 40 |
| SPAN3 | B11CL | 83 | 28 |
| SPAN3 | B11RV | 83 | 27 |
| SPAN3 | B12A | 160 | 41 |
| SPAN3 | B12C | 160 | 74 |
| SPAN3 | B12CL | 160 | 39 |
| SPAN3 | B12RV | 160 | 44 |
| SPAN3 | B13A | 175 | 45 |
| SPAN3 | B13C | 175 | 78 |
| SPAN3 | B13CL | 175 | 46 |
| SPAN3 | B13RV | 175 | 48 |
| SPAN3 | B14A | 383 | 146 |
| SPAN3 | B14C | 383 | 206 |
| SPAN3 | B14CL | 383 | 129 |
| SPAN3 | B14RV | 383 | 133 |
| **Total Responses** | | **27,194** | **15,867** |

## Appendix B Details of Model 1, Model 2 and Model 3

*Model 1 – Binary logistic model describing the level of agreement between examiners (all mark remark data)*

| Effects | Parameter | β | se(β) | p | Prob | Odds |
|---|---|---|---|---|---|---|
| Fixed | Cons | 0.01 | 0.71 | 0.99 | 0.50 | 1.01 |
| | Unit Max | -0.01 | 0.00 | 0.14 | 0.50 | 0.99 |
| | Item Max | -0.09 | 0.05 | 0.07 | 0.48 | 0.91 |
| | Zero is a Band | -0.34 | 0.32 | 0.28 | 0.42 | 0.71 |
| | Number of Bands | 0.22 | 0.28 | 0.43 | 0.55 | 1.25 |
| | Marks per Band | 0.80 | 0.57 | 0.16 | 0.69 | 2.23 |
| | Max Marks in a Band | -0.43 | 0.23 | 0.07 | 0.39 | 0.65 |
| | Min Marks in a Band | -0.08 | 0.21 | 0.70 | 0.48 | 0.92 |
| | Max to Min Ratio | 0.26 | 0.28 | 0.36 | 0.56 | 1.29 |
| | High Bands First | -0.06 | 0.17 | 0.73 | 0.49 | 0.94 |
| | Bands Same Throughout | 0.32 | 0.20 | 0.12 | 0.58 | 1.37 |
| | Optional Questions | 0.00 | 0.06 | 0.99 | 0.50 | 1.00 |
| | AO Grids | 0.29 | 0.13 | **0.03** | 0.57 | 1.34 |
| | Centred Final Mark | -0.76 | 0.07 | **0.00** | 0.32 | 0.47 |
| | Squared Centred Final Mark | 0.87 | 0.25 | **0.00** | 0.71 | 2.40 |
| | Generic Band Descriptions | 0.21 | 0.15 | 0.17 | 0.55 | 1.23 |
| | Marking Position | 0.00 | 0.00 | **0.01** | 0.50 | 1.00 |
| | Standardised Facility | -0.03 | 0.03 | 0.26 | 0.49 | 0.97 |
| | Out of School Hours | -0.12 | 0.03 | **0.00** | 0.47 | 0.89 |
| Random | Unit | 0.02 | 0.01 | 0.06 | 0.51 | 1.02 |
| | Examiner | 0.11 | 0.01 | **0.00** | 0.53 | 1.11 |
| | Item | 0.03 | 0.01 | **0.01** | 0.51 | 1.03 |
| | Response | 1.00 | 0.00 | | | |

*Convergence: RIGLS PQL; Explained Variation $R^2$ = 0.055* (Snijders & Bosker, 1999, p. 225)*; Predictive Efficiency = 0.045* (Long, 1997, p. 106)

$$\text{Agree}_{ijkl} \sim \text{Binomial}(\text{Denom}_{ijkl}, \pi_{ijkl})$$

$$\begin{aligned}\text{logit}(\pi_{ijkl}) = {} & \beta_{0jkl}\text{Cons} + \text{-0.01(0.00)Unit Max}_{l} + \text{-0.09(0.05)Item Max}_{jkl} + \text{-0.34(0.32)Level Zero\_1}_{jkl} + \\ & \text{0.22(0.28)Number of Bands}_{jkl} + \text{0.80(0.57)Marks per Band}_{jkl} + \text{-0.43(0.23)Max Marks per Band}_{jkl} + \\ & \text{-0.08(0.21)Min Marks per Band}_{jkl} + \text{0.26(0.28)Max to Min Ratio}_{jkl} + \text{-0.06(0.17)Low or High\_1}_{l} + \\ & \text{0.32(0.20)Uniform Bands\_1}_{l} + \text{0.00(0.06)Optional or Compulsary\_1}_{jkl} + \text{0.29(0.13)AO Grids\_1}_{jkl} + \\ & \text{-0.76(0.07)Centred Final Mark}_{ijkl} + \text{0.87(0.25)Sqr Centred Final}_{ijkl} + \text{0.21(0.15)Generic\_1}_{l} + \\ & \text{0.00(0.00)Marking Position}_{ijkl} + \text{-0.03(0.03)Standardised Facility}_{jkl} + \text{-0.12(0.03)out of office}_{ijkl}\end{aligned}$$

$$\beta_{0jkl} = 0.01(0.71) + f_{0l} + v_{0kl} + u_{0jkl}$$

*Model 2 – Linear model describing the absolute difference between examiners (all mark remark data)*

| Effects | Parameter | β | se(β) | p |
|---|---|---|---|---|
| Fixed | Cons | 0.02 | 0.80 | 0.98 |
| | Unit Max | 0.00 | 0.00 | 0.91 |
| | Item Max | 0.06 | 0.05 | 0.28 |
| | Zero is a Band | -0.39 | 0.35 | 0.26 |
| | Number of Bands | 0.12 | 0.30 | 0.68 |
| | Marks per Band | 0.41 | 0.60 | 0.49 |
| | Max Marks in a Band | -0.22 | 0.24 | 0.35 |
| | Min Marks in a Band | -0.14 | 0.22 | 0.54 |
| | Max to Min Ratio | -0.12 | 0.28 | 0.68 |
| | High Bands First | 0.22 | 0.18 | 0.21 |
| | Bands Same Throughout | -0.42 | 0.25 | 0.09 |
| | Optional Questions | 0.03 | 0.05 | 0.58 |
| | AO Grids | -0.24 | 0.14 | 0.09 |
| | Centred Final Mark | 0.42 | 0.06 | **0.00** |
| | Squared Centred Final Mark | 0.10 | 0.19 | 0.61 |
| | Generic Band Descriptions | -0.19 | 0.18 | 0.28 |
| | Marking Position | 0.00 | 0.00 | **0.00** |
| | Standardised Facility | 0.03 | 0.02 | 0.22 |
| | Out of School Hours | 0.07 | 0.02 | **0.00** |
| Random | Unit | 0.04 | 0.02 | **0.02** |
| | Examiner | 0.05 | 0.01 | **0.00** |
| | Item | 0.05 | 0.01 | **0.00** |
| | Response | 2.55 | 0.02 | **0.00** |

*Convergence RIGLS; Explained Variation $R^2$ = 11.47%* (Snijders & Bosker, 1999, p. 104)

$$\text{TrueAbsDiff}_{ijkl} \sim N(XB, \Omega)$$

$$\begin{aligned}\text{TrueAbsDiff}_{ijkl} = {} & \beta_{0ijkl}\text{Cons} + \text{-0.00(0.00)Unit Max}_{l} + \text{0.06(0.05)Item Max}_{jkl} + \text{-0.39(0.35)Level Zero\_1}_{jkl} + \\ & \text{0.12(0.30)Number of Bands}_{jkl} + \text{0.41(0.60)Marks per Band}_{jkl} + \text{-0.22(0.24)Max Marks per Band}_{jkl} + \\ & \text{-0.14(0.22)Min Marks per Band}_{jkl} + \text{-0.12(0.28)Max to Min Ratio}_{jkl} + \text{0.22(0.18)Low or High\_1}_{l} + \\ & \text{-0.42(0.25)Uniform Bands\_1}_{l} + \text{0.03(0.05)Optional or Compulsary\_1}_{jkl} + \text{-0.24(0.14)AO Grids\_1}_{jkl} + \\ & \text{0.42(0.06)Centred Final Mark}_{ijkl} + \text{0.10(0.19)Sqr Centred Final}_{ijkl} + \text{-0.19(0.18)Generic\_1}_{l} + \\ & \text{-0.00(0.00)Marking Position}_{ijkl} + \text{0.03(0.02)Standardised Facility}_{jkl} + \text{0.07(0.02)out of office}_{ijkl}\end{aligned}$$

$$\beta_{0ijkl} = 0.02(0.80) + f_{0l} + v_{0kl} + u_{0jkl} + e_{0ijkl}$$

$$-2*\text{loglikelihood(IGLS Deviance)} = 103409.48(27194 \text{ of } 27194 \text{ cases in use})$$

*Model 3 – Linear model describing the absolute difference between examiners (only data where there is a difference)*

| Effects | Parameter | β | se(β) | p |
|---|---|---|---|---|
| Fixed | Cons | 2.37 | 1.48 | 0.11 |
| | Unit Max | -0.01 | 0.01 | 0.48 |
| | Item Max | 0.08 | 0.09 | 0.36 |
| | Zero is a Band | -1.08 | 0.59 | 0.07 |
| | Number of Bands | 0.05 | 0.50 | 0.93 |
| | Marks per Band | 1.07 | 0.96 | 0.27 |
| | Max Marks in a Band | -0.62 | 0.37 | 0.10 |
| | Min Marks in a Band | -0.31 | 0.37 | 0.40 |
| | Max to Min Ratio | -0.06 | 0.45 | 0.89 |
| | High Bands First | 0.45 | 0.32 | 0.16 |
| | Bands Same Throughout | -0.50 | 0.48 | 0.31 |
| | Optional Questions | 0.06 | 0.05 | 0.21 |
| | AO Grids | -0.14 | 0.26 | 0.58 |
| | Centred Final Mark | 0.25 | 0.08 | **0.00** |
| | Squared Centred Final Mark | 2.05 | 0.27 | **0.00** |
| | Generic Band Descriptions | -0.22 | 0.34 | 0.52 |
| | Marking Position | 0.00 | 0.00 | **0.00** |
| | Standardised Facility | -0.01 | 0.02 | 0.79 |
| | Out of School Hours | 0.04 | 0.03 | 0.18 |
| | Proportion Agreeing | -1.55 | 0.53 | **0.00** |
| Random | Unit | 0.18 | 0.07 | **0.01** |
| | Examiner | 0.05 | 0.01 | **0.00** |
| | Item | 0.03 | 0.01 | **0.02** |
| | Response | 2.52 | 0.03 | **0.00** |

*Convergence RIGLS; Explained Variation $R^2$ = 18.64%* (Snijders & Bosker, 1999, p. 104)

$$\text{TrueAbsDiff}_{ijkl} \sim N(XB,\ \Omega)$$

$$\begin{aligned}\text{TrueAbsDiff}_{ijkl} = {} & \beta_{0ijkl}\text{Cons} + \text{-0.01(0.01)Unit Max}_{l} + \text{0.08(0.09)Item Max}_{jkl} + \text{-1.08(0.59)Level Zero\_1}_{jkl} + \\ & \text{0.05(0.50)Number of Bands}_{jkl} + \text{1.07(0.96)Marks per Band}_{jkl} + \text{-0.62(0.37)Max Marks per Band}_{jkl} + \\ & \text{-0.31(0.37)Min Marks per Band}_{jkl} + \text{-0.06(0.45)Max to Min Ratio}_{jkl} + \text{0.45(0.32)Low or High\_1}_{l} + \\ & \text{-0.50(0.48)Uniform Bands\_1}_{l} + \text{0.06(0.05)Optional or Compulsary\_1}_{jkl} + \text{-0.14(0.26)AO Grids\_1}_{jkl} + \\ & \text{0.25(0.08)Centred Final Mark}_{ijkl} + \text{2.05(0.27)Sqr Centred Final}_{ijkl} + \text{-0.22(0.34)Generic\_1}_{l} + \\ & \text{-0.00(0.00)Marking Position}_{ijkl} + \text{-0.01(0.02)Standardised Facility}_{jkl} + \text{0.04(0.03)Out of Hours\_1}_{ijkl} + \\ & \text{-1.55(0.53)Proportion Agreeing}_{jkl}\end{aligned}$$

$$\beta_{0ijkl} = 2.37(1.47) + f_{0l} + v_{0kl} + u_{0jkl} + e_{0ijkl}$$

*-2*loglikelihood(IGLS Deviance)* = 60122.64(15867 of 15867 cases in use)