Levels-based mark schemes and marking bias

Anne Pinot de Moira

Summary

Reliability and validity are ever-present themes in the evaluation of assessment. Previous research studies have considered the role of the question paper, the mark scheme and the examiners in providing unbiased estimates of students' ability. With particular reference to the mark scheme, this paper focuses on how the rules imposed by levels-based mark schemes might affect the extent to which a measurement can be considered unbiased. It considers neither cognitive strategies nor the role of judgement instead, from a purely statistical point of view, it investigates the potential impact of mark scheme design. It concludes that the responses to a question itself should shape the mark distribution, rather than the shape being determined by the structure of the mark scheme. To this end, and pending further research in the area, it recommends that levels-based mark schemes should be designed such that each band is of equal width.

Keywords: Mark schemes, levels-based, banded, marking bias

Introduction

According to classical test theory the observed mark for a candidate is the sum of two independent parts, the true mark for that candidate and measurement error (Lord & Novick, 1968).

Observed Mark = True Mark + Measurement Error

The theory assumes that the error is random and normally distributed. Thus, across a set of candidates, the distribution of observed marks should replicate that of the true marks, assuming that the errors are uncorrelated. A well-designed test should, therefore, provide a set of marks which reflect the true distribution of ability in a given discipline. However, effective measurement relies on effective test development and a test is only as good as its component parts. The reliability and validity of a test can be compromised at any stage of the process. For example, a poorly specified syllabus might affect the validity of the test and, likewise, a badly designed test might not measure what it purports to measure. Poor examiners or ambiguity in the test might preclude reliable marking and affect the extent to which the measurement error can be described as random but, so too, might an ineffective mark scheme. It is the influence of the mark scheme on effective measurement which provides the focus for this paper.

There is a considerable bank of research considering marking, mark scheme design and the implications for reliable marking. Mark remark reliability studies have often been used as a tool to identify features of a mark scheme which might give rise to problems in marking (see for example Baird, Greatorex, & Bell, 2003; Black, Suto, & Bramley, 2011; Massey & Raikes, 2006). Delap (1993) presented the quantitative results of a mark-remark experiment alongside a discussion of the shortcomings of the mark scheme as judged by the participants of the study. That study suggested a redesign to the mark scheme and a subsequent evaluation of the modifications was reported in Baird and Pinot de Moira (1997).

In the more general context of improving the reliability and validity of a test, Moskal and Leydens (2000) recognised the role of the mark scheme. They suggested that clarity might be improved by ensuring that scoring categories are well defined and that the differences between

score categories are clear. These ideas were extended in Pinot de Moira (2011) where the value placed on individual marks was considered. She said that:

> *"each mark must have the same worth in order that the relative weight of items, or assessment objectives, within a paper is as intended. Equally, every mark should be available to be awarded … any item with underutilised marks has the potential to limit discrimination between candidates."* (p. 11)

Pollitt and Ahmed (2008) also recognised the influence that a mark scheme can exert on the demand of a paper; arguing that any performance measure reflects the combined effect of the question and the mark scheme. Accordingly, they defined a valid assessment not only in terms of the paper or question but also in terms of the mark scheme.

> *"An exam task can only contribute to valid assessment: if the students' minds are doing the things we want them to show us they can do; and if we give credit for, and only for, evidence that shows us they can do it."* (p. 2)

Having identified the significance of the mark scheme in ensuring valid assessment, the natural progression is to consider the relationship between the examiner and the mark scheme. Recent research has attempted to capture the cognitive strategies used when marking examination papers. *Think aloud* techniques have been used to distil the features of the judgement process (Crisp, 2008; Suto & Greatorex, 2008a, 2008b). Despite their limitations, think aloud experiments have suggested structure to the thought processes; implying that examiners employ a range of strategies that they modify according to the demands of the question being marked.

However, neither the work on reliability and validity, nor that on cognitive strategy has focused on how the rules imposed upon measurement by the mark scheme might affect the extent to which a test can be reduced to the model proposed by classical test theory. In other words, whether the mark scheme itself might compromise efforts to provide an unbiased assessment by introducing systematic error to the process. Indeed, Lumley (2002) highlighted the fact that "we have no basis for evaluating the judgement that would have been made if a different scale were used" (p. 268).

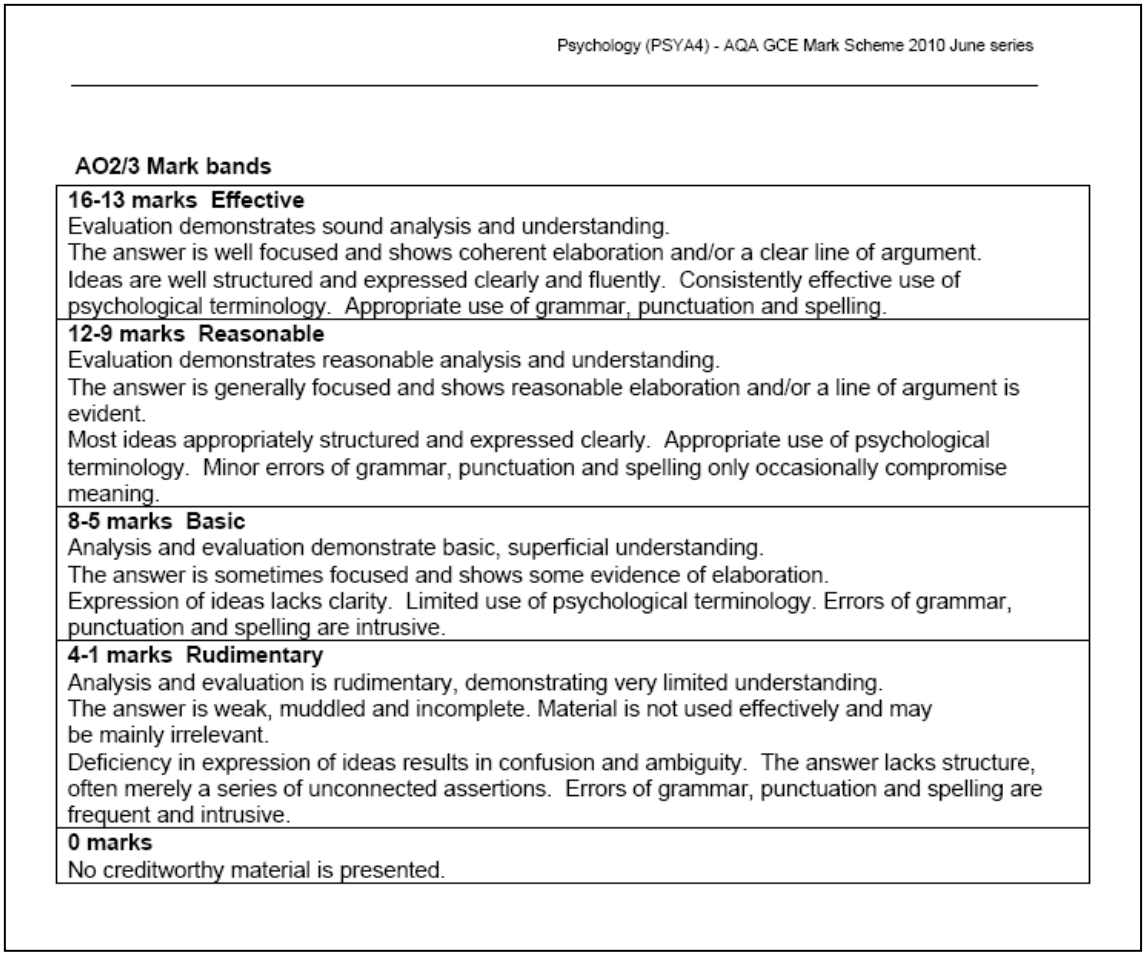
This paper considers the implications to the mark distribution of banded mark schemes or what is generally described as levels-based marking. It does not evaluate the role of judgement, it merely considers the potential effects of rules applied from a simplistic statistical view-point.

Levels-based marking

Levels-based mark schemes are used predominantly for questions with a high mark tariff where there is an extended written response. Such questions have scope for multiple valid approaches, rendering point-based marking or the provision of exemplar answers impractical. Levels-based marking relies on the expert judgement of examiners and has been shown to provide higher reliability than a points-based mark scheme for questions with a maximum mark over six (Bramley, 2008).

In theory, an examiner is required to make an initial assessment of a response and, once the response is classified into a single level, the examiner is then required to refine this judgement to award a single mark within that level. At their simplest, the levels relate to assessment objectives described in the syllabus and the mark scheme gives broad details of the extent to which a response must fulfil the objectives collectively. Figure 1 provides an example of a levels-based mark scheme devised for a question with a maximum tariff of 16. Some levels-

based mark schemes, on the other hand, are more complex. They allow independent judgement of the level for each assessment objective and are therefore structured as a grid.



Psychology (PSYA4) - AQA GCE Mark Scheme 2010 June series

AO2/3 Mark bands

16-13 marks Effective Evaluation demonstrates sound analysis and understanding. The answer is well focused and shows coherent elaboration and/or a clear line of argument. Ideas are well structured and expressed clearly and fluently. Consistently effective use of psychological terminology. Appropriate use of grammar, punctuation and spelling.
12-9 marks Reasonable Evaluation demonstrates reasonable analysis and understanding. The answer is generally focused and shows reasonable elaboration and/or a line of argument is evident. Most ideas appropriately structured and expressed clearly. Appropriate use of psychological terminology. Minor errors of grammar, punctuation and spelling only occasionally compromise meaning.
8-5 marks Basic Analysis and evaluation demonstrate basic, superficial understanding. The answer is sometimes focused and shows some evidence of elaboration. Expression of ideas lacks clarity. Limited use of psychological terminology. Errors of grammar, punctuation and spelling are intrusive.
4-1 marks Rudimentary Analysis and evaluation is rudimentary, demonstrating very limited understanding. The answer is weak, muddled and incomplete. Material is not used effectively and may be mainly irrelevant. Deficiency in expression of ideas results in confusion and ambiguity. The answer lacks structure, often merely a series of unconnected assertions. Errors of grammar, punctuation and spelling are frequent and intrusive.
0 marks No creditworthy material is presented.

Figure 1 Excerpt from the June 2010 AQA A-level Psychology A Unit 4 (PSYA4) mark scheme

The marks within a level are sometimes described individually but, more often, they are merely designed to allow distinction between higher and lower performances within a level. Even though the levels-based mark scheme introduces the element of expert judgement, the mark schemes normally support the decision making with lists of indicative content. Furthermore, the instructions to examiners often describe explicitly the judgement process, thereby restricting the level of expert judgement:

> *"Examiners should initially make a decision about which Level any given response should be placed in. Having determined the appropriate Level the examiners must then choose the precise mark to be given within that Level. In making a decision about a specific mark to award, it is vitally important to think first of the mid-range within the Level …"*
>
> AQA A-level Government & Politics (GOV3C), January 2010 (p. 3)

> *"The first stage is to decide the overall level and then whether the work represents high, mid or low performance within the level."*
>
> Edexcel GCE History (6524 Paper 4F), January 2009 (p. 4)

> *"Tasks that require candidates to respond in extended writing are marked in terms of levels of response. In deciding which level of response to award, teachers should look for the 'best fit' bearing in mind that weakness in one area may be compensated for by strength in another. In deciding which mark within a particular level to award to any response, teachers are expected to use their professional judgement."*
>
> CCEA Learning for Life & Work, 2010 (p. 1)

Whether or not the examiners follow the two-stage decision making process is a moot point but, on the assumption they do, the hierarchical structure of the mark scheme may have implications for the reliability of the marks awarded.

A naïve model of decision making

A hypothetical example

It is possible to model the two-stage decision making process by making distributional assumptions about the decisions at each stage. Without recourse to more qualitative analysis, it is impossible to say whether these assumptions hold true. Nevertheless, it is easy to demonstrate that the design of a levels-based mark scheme could affect the distribution of marks awarded and thus the discrimination between candidates.

For illustrative purposes, assume that the ability distribution of a set of candidates for a given question follows a normal curve and that the levels-based mark scheme for that question follows the simple structure exemplified in Figure 1. The level of a response is determined in accordance with probabilities derived from the normal distribution and, once determined, there is an equal chance that any mark within a level is chosen. Therefore, the choice of mark within a level follows a uniform distribution. Unlike Figure 1, the question in this illustration has a maximum tariff of 14, has five levels and, within each level there are three possible marks that can be awarded[1].

With the given assumptions and parameters, Figure 2 describes the awarded mark distribution and allows comparison with the true mark distribution. The mean mark is denoted by the vertical line and, for this hypothetical example, the true mean and the mean dependent upon the marking instructions are identical. Within the limits of the discrete mark scheme, the observed mark distribution appears to reflect the true mark distribution relatively accurately.

[1] A mark of zero is valid and therefore a question with a maximum tariff of 14 will have 15 possible marks for award.

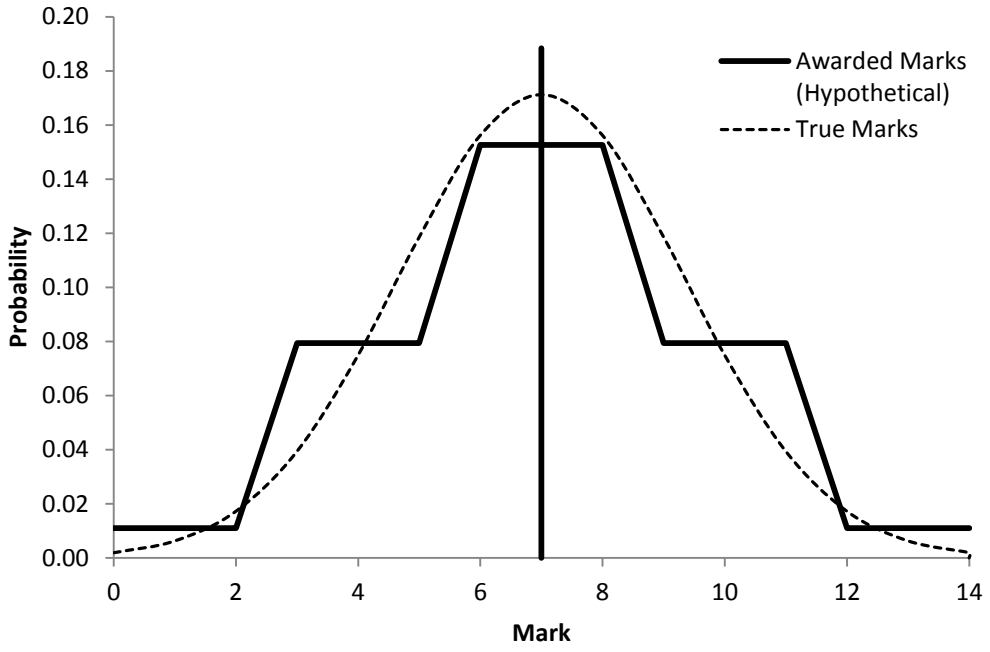


Figure 2 **The distribution of marks awarded for a hypothetical question with maximum mark of 14, five levels and three marks within each level**

A-level English Literature summer 2010

By applying the same assumptions to the two-stage decision making process it is, however, quite possible to show how mark scheme design might affect the distribution of marks that are awarded. This time, the model is applied to a set of questions that appeared on an A-level English Literature paper in summer 2010. The maximum mark tariff was 21. There were 6 levels and the marks were distributed unevenly between these levels. Working from the bottom, there were four marks available in the first level, three in the next two levels and four in the top three. Therefore the number of marks in each level was uneven. The hypothetical distribution of marks awarded would have been as shown in Figure 3. There is a very clear positive skew to the distribution (Skew $_{(hypothetical)}$ = 0.28). The mean awarded mark would be over 5% lower than that suggested by the true mark distribution. Thus more candidates would be clustered towards the bottom of the distribution than should be, given their underlying ability.

Figure 3 also includes the distribution of marks awarded to candidates in the live examination. This distribution is even more skewed than that suggested by the model (Skew $_{(live)}$ = 0.59) and so, unsurprisingly given its arbitrary nature, the model fails to describe fully the mechanisms at work in the examiner decision making process. It is noteworthy, however, that the skewed mark distribution for this A-level unit provided cause for concern and was investigated in the context of improving marking reliability. Whatever the mechanisms underlying the reluctance to award marks towards the top of the mark distribution, it is surely possible that they were exacerbated by the asymmetric levels in the mark scheme.

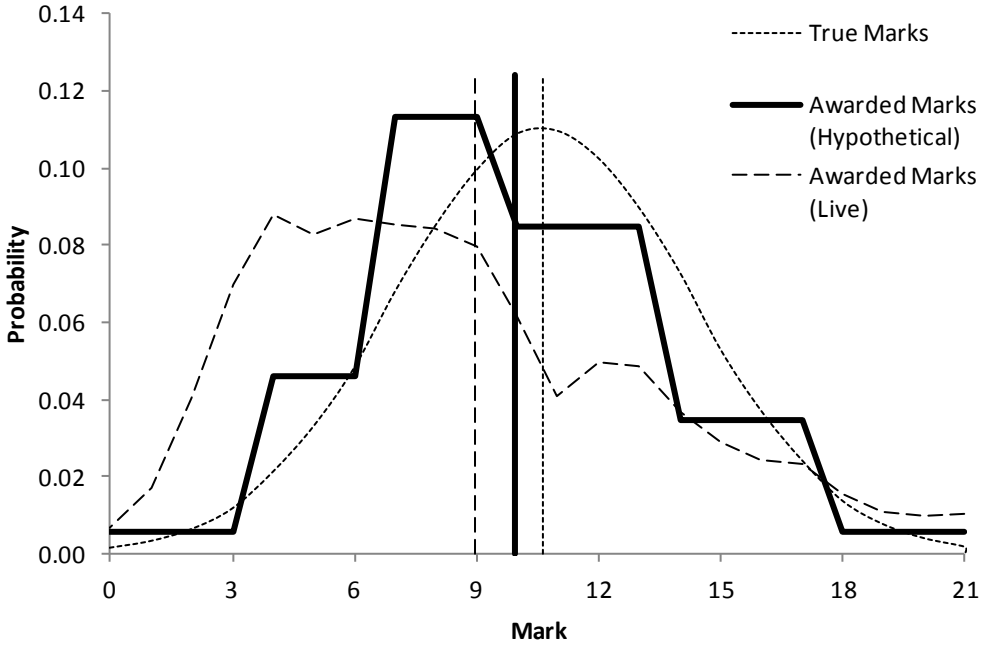


Figure 3 **The distribution of marks awarded for an question with maximum mark of 21, six levels and differing marks within each level**

Other levels-based mark schemes

The appendices give further examples of the potential effect of the structure imposed by levels-based mark schemes. The illustrations include examples of the effect of changes to the mark band structure (Appendix A, Figures A.1 to A.6) and changes to the underlying distribution of ability (Appendix B, Figures B.1 to B.6). In the latter group, the Weibull distribution has been used as a substitute for the normal distribution to allow modelling of a skew in the ability of the entry. Evidence suggests that no matter what the underlying distribution of ability, the best estimate of the true mean mark for a cohort of candidates is given when each mark band is of equal width.

Limitations

The model presented here is naïve because it defines the decision making process in terms of arbitrarily assigned statistical distributions. If the distributional assumptions were changed, the findings would change. Furthermore, if the instructions for levels-based marking were modified, personalised or ignored by examiners, any efforts to model statistically the decision making process would surely be in vain. Indeed, Lumley (2002) argues that it is the examiner, rather that the mark scheme, that lies at the centre of the marking process.

Nevertheless, as they stand, the rules laid out in many mark schemes recommend a two-stage judgement and by way of confirmation, in her qualitative investigation of the judgement process in examination marking, Crisp (2010) conceptualises decision making as a multi-stage process. Her think-aloud verbalisations indicate that a final mark is only awarded after an initial evaluation of the strengths and weaknesses of the response where a level is determined.

> *"… there seems to be a somewhat automatic production of an approximate level of the response thus narrowing the mark range …"*
>
> (p.15)

Both Crisp's work, and that of Suto and Greatorex (2008a), implicitly suggests some interaction with the mark scheme in the judgement process. Therefore, it must be worth considering the possibility that the way in which marks are distributed between levels of response has the potential to bias the awarded marks distribution.

The mark for an individual question, however, is never used in isolation. It is invariably combined with the marks from other questions to produce an estimate of ability. Each question on an examination paper will have its own mark scheme and the multiplicative effect of each mark scheme, biased or otherwise, is more complex than the model presented here. Nevertheless with question level marking, which suggests independence between the marks awarded for different questions on a unit, the central limit theorem[2] might lead one to conclude that the mark distribution on individual units is irrelevant as long as the distribution of marks for each question can be said to be identical. In other words, whatever the distribution of marks awarded for individual questions, the distribution of marks on the unit as a whole should be normally distributed. This, of course, would only be a desirable property if the underlying distribution of ability was normal and, even if that were the case, three important issues arise.

Firstly, given the rules imposed by the mark scheme, it is unlikely that judgements made for each separate question are independent. The mark distribution for the entire A-level English Literature unit, for example, was skewed even after all questions were aggregated to produce a unit outcome (Skew $_{(unit)}$ = 0.56[3]). Secondly, given the bespoke nature of mark schemes, with differing instructions for each question, it is unlikely that the distribution of marks for each question would be identical. Thirdly, and perhaps most importantly, even if the unit mark distribution appeared superficially to replicate the true mark distribution, the fact that the contributing mark distributions were skewed might affect the rank order of candidates. Therefore, over and above the random measurement error, there would be an unintended and systematic element to the error due to the structure of the mark scheme.

It could be argued that the distribution of marks for a given question might be intentionally skewed even if the ability of candidates across the whole unit follows a normal distribution. Questions assessing more complex areas of the syllabus might elicit few excellent responses and a larger number of mediocre or poor responses. Each question represents a slice of the syllabus and different areas of the syllabus might give rise to different patterns of achievement. In that case, different distributional assumptions would be needed in the model. However, the responses themselves should generate the correct mark distribution, rather than the shape being determined by the structure of the mark scheme.

It is interesting to reflect that, in the context of the equitable award of grade A* in the A-level qualification, Pinot de Moira (2007) highlighted the same issue. She suggested that the shape of the mark distribution should reflect only the characteristics of the entry rather than any bias in the assessment or mark scheme.

Discussion and conclusions

Returning briefly to classical test theory, a candidate's true mark is said to be given by the pooled judgement of an infinite number of judges (Wiseman, 1949). The theory is only justifiable if the error associated with measurement is random. Assessment is an imprecise

[2] The central limit theorem states that as the number of independent, identically distributed random variables with finite variance increases, the distribution of their mean becomes increasingly normal.

[3] The positive skew for the unit outcome was in contrast to a very marginal negative skew in candidates' prior attainment as measured by mean GCSE score (Skew $_{(mean\ GCSE)}$ = -0.10).

science and therefore error can be introduced at any stage of the process and, within reason, so long as the error is unbiased and small it may be considered acceptable.

A mark scheme is designed to reduce error in marking and provide a structure to the judgement process. Dependent upon the nature of the question, the structure might be rigid or might simply provide a scaffold for expert judgement. Regardless, it should be designed such that the marks awarded are unbiased and there is no systematic error associated with the marking process.

Even the naïve model presented here suggests that the structure of a mark scheme could have implications for the marks awarded. The variety of levels-based mark schemes currently used in the national testing arena suggests that design implications are not a preeminent concern in their development. In conjunction with the qualitative research considering the judgement process, improvements in the reliability and validity of marks awarded might be made by considering the structure of levels-based mark schemes. In the meantime, in the absence of any clear educational imperative to the contrary, a conservative approach to the design of such mark schemes would be to ensure that the band widths are designed to be of equal width.

References

Baird, J., Greatorex, J., & Bell, J. (2003). What makes marking reliable? Experiments with UK examinations. *AQA Internal Report*, *RPA_03_JB_RC_217*.

Baird, J., & Pinot de Moira, A. (1997). Marking reliability in Summer 1996 A Level Business Studies. *AQA Internal Report*, *RPA_97_JB_RAC_760*.

Black, B., Suto, I., & Bramley, T. (2011). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement. *Assessment in Education: Principles, Policy and Practice*, *18*(3), 295-318.

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

Crisp, V. (2008). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education*, *38*(2), 247-264.

Crisp, V. (2010). Towards a model of the judgement processes involved in examination marking. *Oxford Review of Education*, *36*(1), 1-21.

Delap, M. (1993). Marking Reliability Study In Business Studies (0655). A Study Of The June 1992 Examination, Papers 1 And 2. *AQA Internal Report*, *RPA_93_MD_RAC_609*.

Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, Massachusetts: Addison-Wesley Publishing Company, Inc.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*(3), 246-276.

Massey, A., & Raikes, N. (2006). Item-level examiner agreement. In *British Educational Research Association, Annual Conference*. Warwick, UK.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*, *7*(10), Retrieved March 1, 2011 from http://PAREonline.net/getvn.asp?v=7&n=10.

Pinot de Moira, A. (2011). Effective discrimination in mark schemes. *AQA Internal Report*, *CERP_11_APM_RP_013*.

Pinot de Moira, A. (2007). An effort to cap it all: The cap and its effect on the award of grade A* at A Level. *AQA Internal Report*, *RPA_07_APM_TR_015*.

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

Suto, I., & Greatorex, J. (2008a). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policy & Practice*, *15*(1), 73-89.

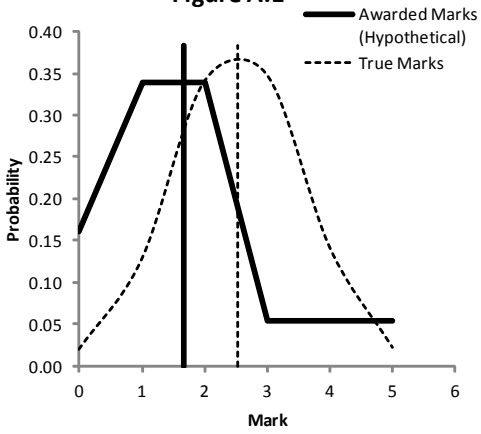
Suto, I., & Greatorex, J. (2008b). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*, *34*(2), 213-233.

Wiseman, S. (1949). The marking of English composition in grammar school selection. *British Journal of Educational Psychology*, *19*(3), 200-209.

24 November 2011

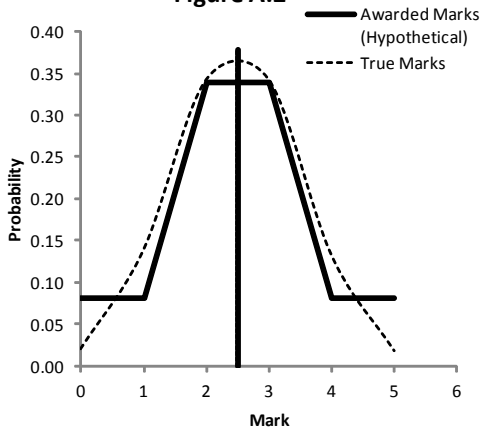
Appendix A Other Levels-Based Mark Schemes (Normal Distribution)

Figure A.1



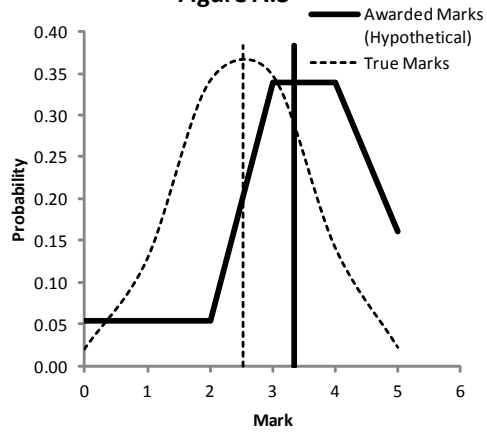
Marks in each band:	1, 2, 3
Difference in Mean:	-31.90%
Skew in Awarded Marks:	0.97

Figure A.2



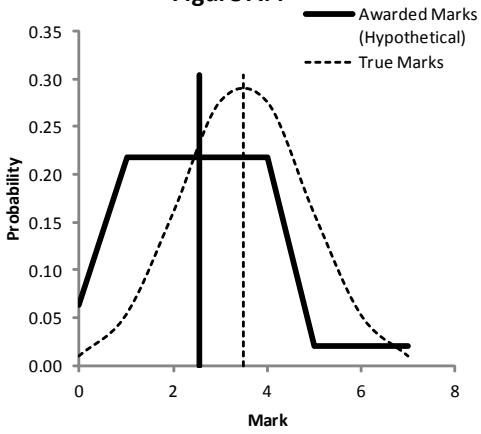
Marks in each band:	2, 2, 2
Difference in Mean:	0.00%
Skew in Awarded Marks:	0.00

Figure A.3



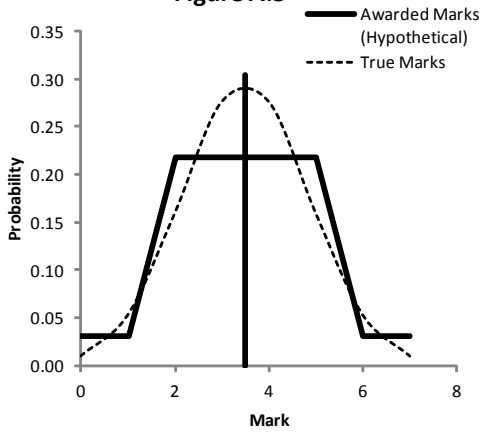
Marks in each band:	3, 2, 1
Difference in Mean:	31.90%
Skew in Awarded Marks:	-0.97

Figure A.4



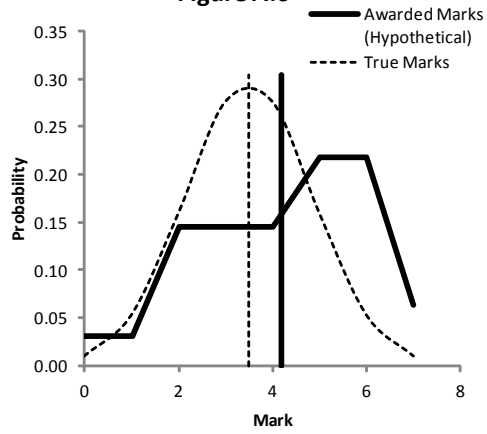
Marks in each band:	1, 2, 2, 3
Difference in Mean:	-26.76%
Skew in Awarded Marks:	0.50

Figure A.5



Marks in each band:	2, 2, 2, 2
Difference in Mean:	0.00%
Skew in Awarded Marks:	0.00

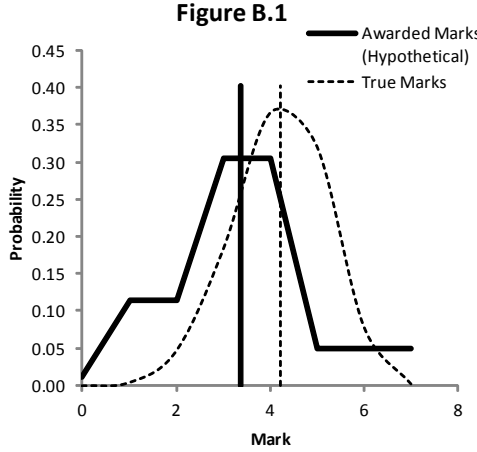
Figure A.6



Marks in each band:	2, 3, 2, 1
Difference in Mean:	19.62%
Skew in Awarded Marks:	-0.41

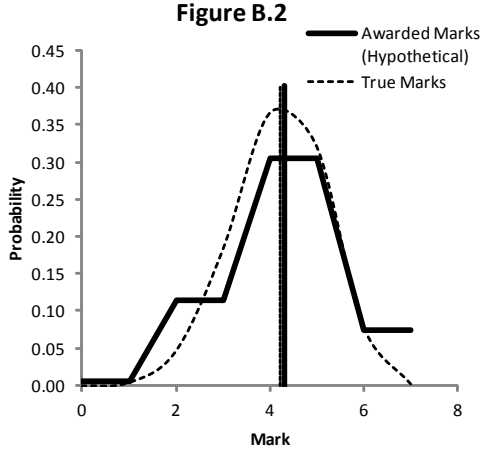
Appendix B Other Levels-Based Mark Schemes (Weibull Distribution)

Figure B.1



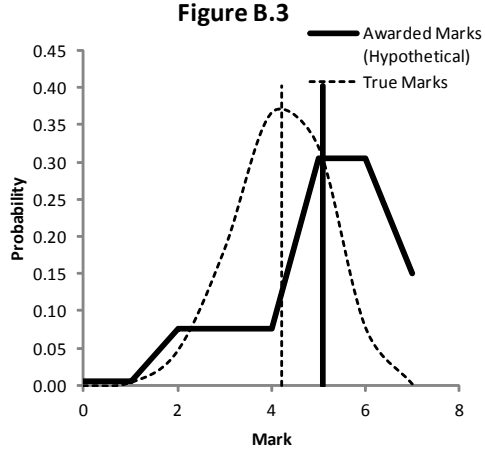
Marks in each band:	1, 2, 2, 3
Difference in Mean:	-19.62%
Skew in Awarded Marks:	0.37
Skew in True Marks:	-0.25
Weibull Parameters:	β=5, η=2

Figure B.2



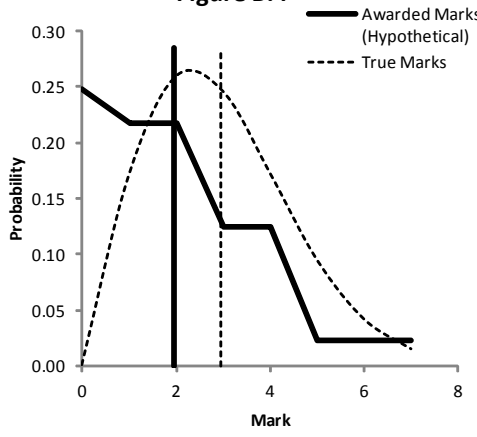
Marks in each band:	2, 2, 2, 2
Difference in Mean:	2.26%
Skew in Awarded Marks:	-0.13
Skew in True Marks:	-0.25
Weibull Parameters:	β=5, η=2

Figure B.3



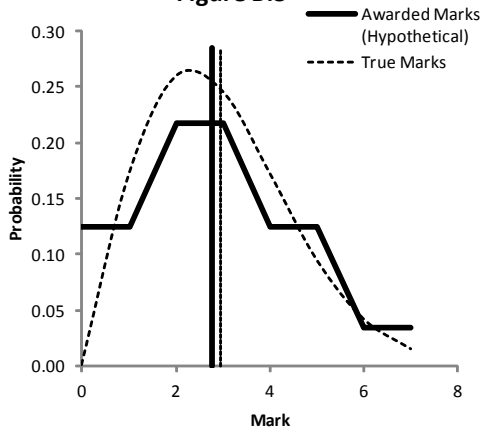
Marks in each band:	2, 3, 2, 1
Difference in Mean:	21.29%
Skew in Awarded Marks:	-0.90
Skew in True Marks:	-0.25
Weibull Parameters:	β=5, η=2

Figure B.4



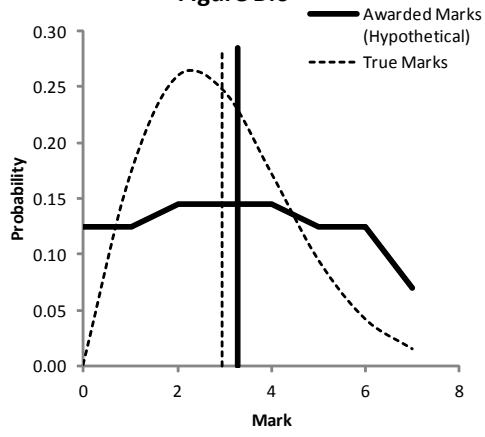
Marks in each band:	1, 2, 2, 3
Difference in Mean:	-34.18%
Skew in Awarded Marks:	0.87
Skew in True Marks:	0.56
Weibull Parameters:	β=2, η=5

Figure B.5



Marks in each band:	2, 2, 2, 2
Difference in Mean:	5.58%
Skew in Awarded Marks:	0.32
Skew in True Marks:	0.56
Weibull Parameters:	β=2, η=5

Figure B.6



Marks in each band:	2, 3, 2, 1
Difference in Mean:	11.41%
Skew in Awarded Marks:	0.07
Skew in True Marks:	0.56
Weibull Parameters:	β=2, η=5