# EFFECTIVE DISCRIMINATION IN MARK SCHEMES

Anne Pinot de Moira


## SUMMARY

Effective assessment is often measured in terms of mark re-mark reliability: the extent to which the same mark would be awarded to a candidate by two different examiners or, alternatively, by the same examiner on two different occasions. Reliability, however, can come at the expense of validity: the extent to which an assessment tests the skills it purports to test. The validity of an assessment can be compromised by any element of the specification, the examination paper or the mark scheme. With the use of a hypothetical illustration and a case study, this paper explores the effect that a mark scheme can have on the validity of an assessment; focussing particularly on a mark's worth across the mark range for an item. It proposes a number of techniques which might be used to identify areas of weakness in a mark scheme and areas where the mark scheme does not provide for effective discrimination between candidates.


*Keywords: Mark scheme, discrimination, partial credit model, disordered thresholds*

## INTRODUCTION

A well-written mark scheme should allow effective discrimination between different responses to an item or question. As with all stages of an assessment process, the aim should be to minimise the error introduced. In the research community, the efficacy of a mark scheme has most often been judged in terms of mark re-mark reliability. As part of a review of the marking reliability literature, Meadows and Billington (2006) discussed the effect of mark schemes on marking reliability. Many of the studies therein referred to consistency between judges or examiners as the key determinant of success when modifying and improving mark schemes.

Experimental findings have defined reliability as, amongst other things, the correlation between the marks awarded by examiners, level of examiner agreement and the magnitude of the difference between the marks awarded by examiners. Bramley (2008) used the proportion of cases with exact agreement to look at the features of a mark scheme which might increase reliability. The statistics reported in these studies have been used to represent the extent to which the same mark would be awarded to a candidate by two different examiners, or by the same examiner on two different occasions. More recently, Ofqual's reliability programme has encouraged the awarding bodies of England & Wales to generate and publish reliability statistics for all their assessments (Opposs & He, 2011) although, to this end, it is questionable whether the statistics suggested would actually provide any meaningful measure of mark-remark reliability.

That a mark scheme produces replicable outcomes when being used under differing circumstances is clearly desirable. However, in focussing solely on this property, one may lose sight of the need to award valid marks to each candidate and to discriminate fairly between these candidates. As Moskal and Leydens (2000, Concluding Remarks, para. 1) highlight, "although a valid assessment is by necessity reliable, the contrary is not true. A reliable assessment is not necessarily valid." To be fair to those engaged in assessing mark re-mark reliability, many concede that validity has an equally important part to play in assessment (see for example, Baird & Pinot de Moira, 1997; Bramley, 2008). Validity, though, is often described

in terms of the content and design of the paper, rather than the design of the mark scheme. However, badly designed mark schemes can easily impair the validity of the assessment when the paper itself may be quite reasonable. On the basis of a qualitative evaluation, Pollitt, Ahmed, Baird, Tognolini, & Davidson (2008) conclude:

> *"It is not enough to write good exam questions … validity principle demands that we also give credit to, and only to, the evidence that [candidates] can do these things." (p. 47)*

In the context of low tariff items, Bramley (2001) argued that an emphasis on content validity in the design of a mark scheme sometimes serves to undermine the statistical validity of an assessment. This paper uses a quantitative approach to extend these arguments and to propose a mechanism for identifying weaknesses in a mark scheme.

## AN ILLUSTRATION

It is simple to construct an example of a short answer paper which, with relatively objective answers, provides high mark re-mark reliability. Imagine an examination paper comprising two items, each of which is marked out of two. For the first item, the mark scheme specifies the responses required to be awarded either zero, one or two marks. For the second item, the mark scheme only specifies the responses required to be awarded zero or two marks. In this illustrative case, the implication is that no response to Item 2 is worthy of one mark alone (see Pollitt et al., 2008, p. 42 for a qualitative description of this scenario). Hence, the range of possible outcomes for any given script is between zero and four (Table 1) and, although all marks in this range are available, candidates are more likely to be awarded two than any other mark. But do the marks awarded at a script level reflect the relative merit of the candidates? It all hinges on whether a mark's worth is equivalent for each item, whether the relative weight of each item is correct and, furthermore, whether there is truly no response worthy of one mark for the second item.

The issue of the value of a mark must underpin the design of any mark scheme in order that the relative weight of items, or assessment objectives, within a paper is as intended. That is not to say that every mark must be explicitly described within the mark scheme, rather that every mark should be available and that each should be of equal merit.

In the context of this illustration, what if the mark of two is not equivalent for both items and the maximum mark for Item 2 has been inflated in an attempt to equate assessment objectives? In other words, Item 2 requires a one word, one mark response which could genuinely be considered completely right or completely wrong. Thus the true merit of the response is one mark and this, in turn, equates to a one mark response on Item 1. It means that the mark scheme, as designed, explicitly distorts the contribution of each assessment objective. Therefore, Item 2, where a mark has greater worth, contributes more than it should to the final assessment in terms of relative weight.

Revising the mark scheme by, for example, setting a maximum mark of one for Item 2, shows the effect that a seemingly trivial feature can have upon the rank order of candidates (Table 1). Within the limitations of this simplistic illustration, it can be seen that a candidate achieving Outcome 3 (zero on the first item and maximum marks on the second) would be ranked above halfway with the original mark scheme, but below halfway on the revised scheme.

On the other hand, if a mark of two is designed to be equivalent for both items, then why is it not possible to award a mark of one for responses to Item 2? For this second item, candidates are unable to show a range of responses because, according to the mark scheme, what is written is

either completely right or completely wrong: there is no gradient. Thus, discrimination between candidates is limited and any design assumption of equally weighted assessment objectives is erroneous. Empirically, the achieved weight might be as intended, but the designed differences in mark scheme for the two items affect candidates differentially dependent upon their area of expertise.

**Table 1 The range of possible outcomes in the illustrative two item paper**

| | Original mark scheme | | | | Revised mark scheme | | | | Same |
|---|---|---|---|---|---|---|---|---|---|
| Outcome | Item 1 | Item 2 | Total | Rank | Item 1 | Item 2 | Total | Rank | rank? |
| 1 | 0 | 0 | 0 | 1.0 | 0 | 0 | 0 | 1.0 | ✓ |
| 2 | 1 | 0 | 1 | 2.0 | 1 | 0 | 1 | 2.5 | × |
| 3 | 0 | 2 | 2 | 3.5 | 0 | 1 | 1 | 2.5 | × |
| 4 | 2 | 0 | 2 | 3.5 | 2 | 0 | 2 | 4.5 | × |
| 5 | 1 | 2 | 3 | 5.0 | 1 | 1 | 2 | 4.5 | × |
| 6 | 2 | 2 | 4 | 6.0 | 2 | 1 | 3 | 6.0 | ✓ |

This illustration can be extended to consider two candidates taking the same paper. Assume that, according to their individual abilities, the first candidate should be awarded one mark on Item 1 and two marks on Item 2 (Table 2), whereas the second candidate should be awarded two marks on Item 1 and one mark on Item 2. All other things being equal, both candidates should be awarded the same mark for the script. However, the second candidate will not receive one mark for Item 2 because the mark scheme implies that no such award can be made. Therefore, dependent the examiner's compromise judgement, Candidate 2 would either be ranked above or below, but not equivalent to, Candidate 1 (Table 2).

**Table 2 The fate of two candidates**

| | | Item 1 | Item 2 | Total |
|---|---|---|---|---|
| Candidate 1 | Correct decision | 1 | 2 | 3 |
| Candidate 2 | Correct decision (not available) | 2 | 1 | 3 |
| | Unfavourable decision | 2 | 0 | 2 |
| | Favourable decision | 2 | 2 | 4 |

Whichever perspective is taken – non equivalent or equivalent mark's worth – there are flaws in the design of the mark scheme described in this illustration. With different mark's worth, the balance of assessment objectives within the paper is likely to be distorted. With mark equivalence but some unavailable marks, the rank of candidates will be directly affected by the pragmatic, rather than correct, decisions made by examiners.

Operationally, there are few such flagrant examples of mark schemes which distort the weighting of items and the rank order of candidates although, regrettably, there are some. Yet within most question papers examples can be found where marks are underutilised and, as shown above, any item with underutilised marks has the potential to influence discrimination between candidates.

This paper makes a first attempt to propose a method for identifying areas of weakness in a mark scheme and assessment by considering a unit that was awarded in July 2010. It focuses on the importance of discriminating effectively between candidates, rather than on the more commonly documented characteristic (see Meadows & Billington, 2006 for examples) of mark re-mark reliability.

## A CASE STUDY

The unit under consideration is the first AS unit from Specification X (Table 3), which has 24 items grouped into eight topics. The topics are spread across two sections. A candidate must choose three topic areas in total and at least one must be from each section. Within each topic, the candidate is required to answer all three items. The first two items in each section are marked out of 10 and the third is marked out of 12; two of the marks for the latter being awarded for quality of written communication[1].

**Table 3 The structure of Unit 1 for Specification X in June 2010**

| Section A | | | Section B | | |
|---|---|---|---|---|---|
| Topic | Item | Marks | Topic | Item | Marks |
| Topic 1 | 01 | 10 | Topic 5 | 13 | 10 |
| | 02 | 10 | | 14 | 10 |
| | 03 | 12 | | 15 | 12 |
| Topic 2 | 04 | 10 | Topic 6 | 16 | 10 |
| | 05 | 10 | | 17 | 10 |
| | 06 | 12 | | 18 | 12 |
| Topic 3 | 07 | 10 | Topic 7 | 19 | 10 |
| | 08 | 10 | | 20 | 10 |
| | 09 | 12 | | 21 | 12 |
| Topic 4 | 10 | 10 | Topic 8 | 22 | 10 |
| | 11 | 10 | | 23 | 10 |
| | 12 | 12 | | 24 | 12 |

### Simple Statistics

It is impossible to know how the true mark distribution for a particular item should look. It is dependent upon the cohort of candidates responding, the effective design of assessment and mark scheme, and the influence of external pressures such as quality control measures. Thus, to assess whether marks have been awarded as would reasonably be expected, a simple strategy might be to compare the distribution of marks with features of the mark scheme.

Consider Item 2 on Unit 1 of Specification X. The distribution of marks for this item is illustrated by the solid line in Figure 1. Relative to the adjacent marks, the proportion of candidates who were awarded six marks appears low. Cross-referencing with the mark scheme throws some light on why this might be the case. In its introductory section, the mark scheme for the unit as a whole gives clear definitions of the terms which can be applied generically throughout all

[1] The examples presented as part of the case study have been chosen to illustrate particular points, so it should not be inferred that problems are endemic within the unit.

items on the paper. Following on from this, there are specific details to aid the marking of each individual item. First, there is a discussion of potential content followed by a description of the mark bands for the item. The mark bands for Item 2 are illustrated in Figure 2 and, for this item, there are two areas of potential content, (A) and (B); not be confused with section A and section B on the paper. In answer to the question, candidates are required to provide an outline of each of these areas of content. Competency in this task is measured against the descriptors provided by the mark bands.

While it is not essential that every mark on a mark scheme is explicitly defined, it is interesting to note that there was only *one* mark on this mark scheme that was *not* defined, and that mark was six. It may be unsurprising, therefore, that this mark appears underutilised.

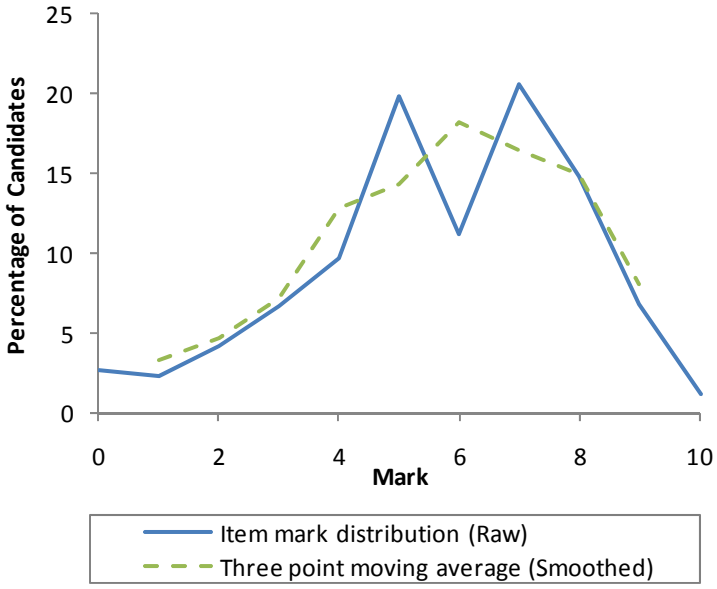


**Figure 1 The mark distribution for Item 2 on Unit 1 of Specification X (Raw and Smoothed)**

| **Mark Band** | |
|---|---|
| 8 – 10 | The candidate deals with (A) and (B) as follows:<br>**max 10:** two sound<br>**max 9:** one sound, one clear<br>**max 8:** one sound, one some or two clear. |
| 5 – 7 | The candidate deals with (A) and (B) as follows:<br>**max 7:** one sound or one clear, one some<br>**max 5:** one clear or two some. |
| 3 – 4 | The candidate demonstrates some understanding of (A) or (B), or limited understanding of (A) and (B). |
| 1 – 2 | The answer consists of brief, fragmented comments or examples so that no coherent explanation emerges<br>**or**<br>mistakes and confusion fundamentally undermine a more substantial attempt at explanation. |
| 0 | The answer contains no relevant information. |

**Figure 2 Mark bands in the mark scheme for Item 2 on Unit 1 of Specification X**

On a case by case basis, it may be possible to eliminate idiosyncrasies in an individual mark scheme, but is it possible to mechanise the process to limit the amount of complex cross-referencing and focus only on problem areas? To do this requires confronting the notion that it is not possible to know how the true mark distribution for a particular item should look. One could assume that performance on any given item should be some roughly smooth bell shaped curve with fewer candidates at the lower and upper ends of some notional ability range and more clustered around the middle. Specifically, maybe the proportion of candidates awarded a given mark could be said to be similar to the proportion awarded the adjacent marks. Such an assumption would probably eliminate any multimodality. Figure 1 on the previous page shows a centred 3-point moving average, calculated as the mean proportion of candidates over the given mark and each of the adjacent marks. The dotted line shows this smoothed mark distribution. Comparing the raw and smoothed mark distribution suggests that the marks of five and seven might be slightly over-utilised and the mark of six slightly underutilised.

These data could also be presented as the difference between the raw mark distribution and the expected mark distribution described by the 3-point moving average. Therefore, if the distribution of differences, or residuals, is approximately normal, standardising them to have a mean of zero and a standard deviation of one, allows comparisons to a normal distribution. The standardised residuals allow identification of unusual observations. Commonly, standardised residuals greater than ± 2 might be considered to misfit the model where the model, in this case, is the smoothed mark distribution. Figure 3 shows the standardised residuals for Item 2 on Unit 1 of Specification X and, for information, the mark bands have been superimposed on the graph[2]. The pattern is very similar to that identified from Figure 1, albeit somewhat starker. Given the expectation of a smoothed mark distribution, the award of five marks is higher than might be expected and the award of six marks is lower than expected.

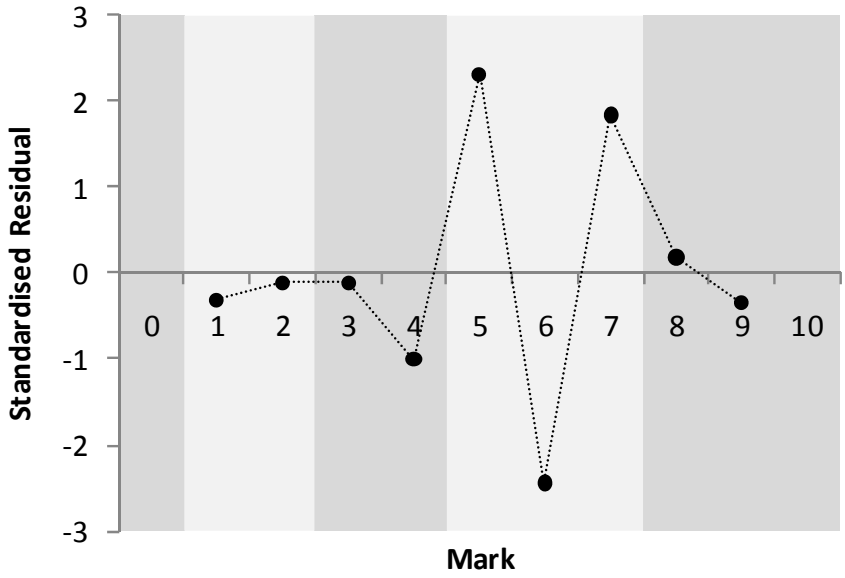


**Figure 3 Standardised residual difference between the raw and smoothed mark distribution for Item 2 on Unit 1 of Specification X**

[2] Note that because it is not possible to create a centred three point moving average for the marks of 0 and 10, residuals for these marks can not be calculated.

The disadvantage of standardising the residuals of individual items is that it renders comparisons between items meaningless. By standardising, each item is engineered to have the same spread of residuals and therefore there is no facility to differentiate between items which fit the model well and items which do not. However, to make comparisons between items of the same tariff it would be possible to standardise the residuals across those items. After accounting for the different entry size between items, all residuals could be adjusted by subtracting the mean residual across all items and dividing by the standard deviation of residuals across all items.

To this end, Figure 4 illustrates the residuals for all 16 of the 10 mark tariff questions on Unit 1 of Specification X. The items have been grouped according to the type of mark scheme. Those with the relatively prescriptive mark scheme shown in Figure 2 are denoted by the line graphs with crosses at the plot points. Those with a more generic, but still banded mark scheme are denoted by the line graphs with the dots at the plot points. There is a clear pattern: the items marked using the prescriptive mark scheme, which omits any description of a mark of 6, deviate more greatly from the smoothed mark distribution than those marked using the generic mark scheme. This is not necessarily to say that the prescriptive mark scheme is at fault, more that the marks awarded using this mark scheme do not give rise to a smooth mark distribution. It is a much more impenetrable exercise to determine whether the model is correct and therefore whether the mark scheme is providing appropriate discrimination to the candidates.

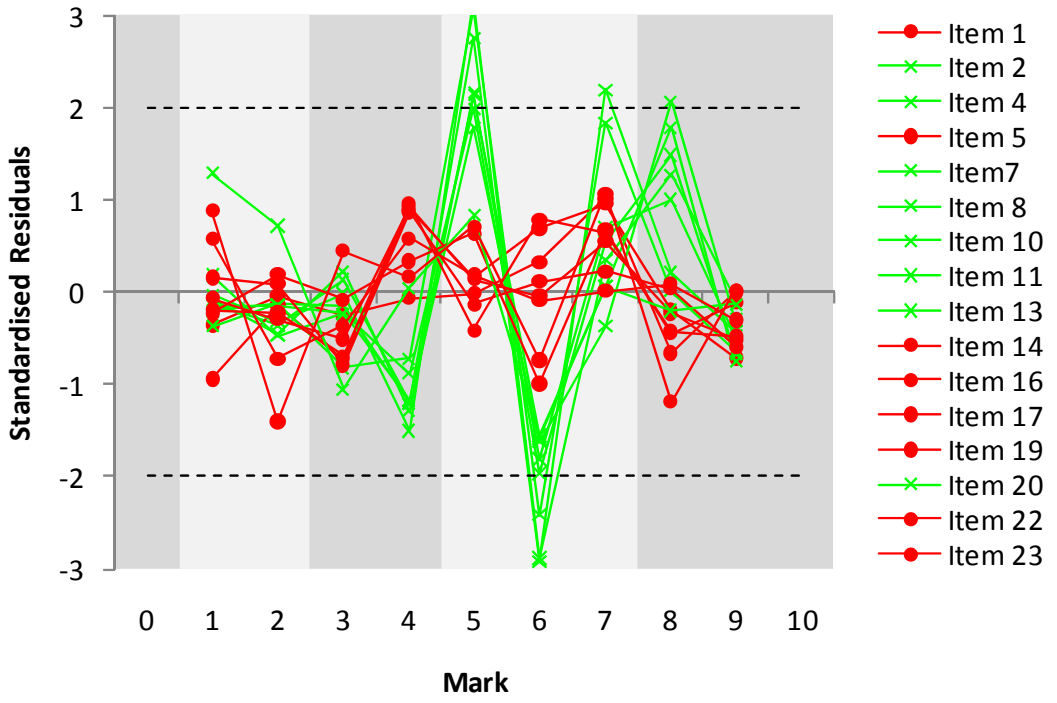


**Figure 4 Standardised residual difference between the raw and smoothed mark distribution for all 10 mark items on Unit 1 of Specification X**

## A Partial Credit Model

*Background*

Notwithstanding the limitations of using a simple moving average as a model of the ideal mark distribution, such a model takes no account of candidature or of question difficulty. It might be a

mark is underutilised simply because there are few candidate entries at that point on the ability range. The potential to extend the method to make comparisons between units, or optional items within units, would also be limited. In a Partial Credit Model (PCM) (Masters, 1982), however, the probability of a specified mark being awarded is modeled as a function of candidate ability and item difficulty, thereby ameliorating some shortcomings of the simple model. For papers with optional questions there remains the limitation that systematic differences in question choice between defined sub groups of the entry could undermine conclusions on relative question difficulty. Nevertheless, in the context of assessing the adequacy of the mark scheme, the evaluation of relative question difficulty is not the primary objective; it is the pattern of marks allocated for each item that is of greater interest.

The PCM provides Rasch-Andrich threshold measures (Andrich, 1978) which quantify the point along the latent continuum at which it is equally likely that either of two adjacent marks will be awarded. Thus, for an item marked out of *n*, there will be *n-1* threshold measures. The distance, in terms of logits, between adjacent thresholds indicates the extent of the continuum over which that category is the modal mark (assuming there are no disordered thresholds). Evenly spread thresholds over a sensible range might indicate an item which discriminates in an unbiased manner. Conversely, unevenly spread thresholds might indicate that there are some marks which are rarely awarded. Indeed, it may be the case that a PCM uncovers disordered thresholds, where a particular mark is never the most likely to be awarded whatever point on the latent continuum.

Disordered thresholds are thought by some to undermine the assumptions underlying the PCM. They may arise from multidimensional, rather than unidimensional, responses and may indicate that the assumption of a single measureable latent trait is unsupported (Andrich, de Jong, & Sheridan, 1997). That said, they also identify areas of the mark scheme where an increase in the assumed single latent trait does not coincide with an equal increase in the credit awarded. Indeed Andrich (1998) observes that data which produce disordered thresholds demonstrate that ordering has not worked as intended. He also suggests that the data may be disordered for a good reason and that the reason should be acknowledged. Linacre (2002, p. 10) summarises the implications of disordered thresholds thus:

> *"Disordering reflects the low probability of observance of certain categories because of the manner in which those categories are being used in the rating process. Thus, … disordering degrades the interpretability of the resulting measures. It can indicate that a category represents too narrow a segment of the latent variable or corresponds to a concept that is poorly defined in the minds of the respondents. Disordering of step calibrations often occurs when the frequencies of category usage follow an irregular pattern."*

These implications can be translated to the context of marking validity. If a PCM reveals disordered thresholds this might imply that, across the mark range for an item, the mark's worth is not equivalent. Thus, because the contribution of different aspects of the specification differs from that intended, the legitimacy of summing item marks might be undermined. Furthermore, interpretation of the resultant summed marks might, at best, be problematical and, at worst, be meaningless. The disordering might also reflect a poorly designed mark scheme and an irregular allocation of marks.

As far as evaluating a mark scheme is concerned, therefore, disordered thresholds provide a good starting point for investigation. Using the same unit from Specification X, a PCM was fitted to explore patterns of mark allocation and shortfall in the mark scheme.

*The Case Study*

Table 4 shows the item difficulty measures derived from the PCM for Specification X. For all items, with the possible exception of Item 5, the model seems to provide a good fit for the data. Assuming no systematic difference between the candidature for each item, the small range of difficulty measures suggests there was little bias in outcome as a result of candidates' option choice.

**Table 4 Item difficulty statistics for Unit 1 of Specification X**

| Topic | Item | Measure | Error | Entry | Infit* | Outfit* |
|---|---|---|---|---|---|---|
| 1 | 1 | -0.27 | 0.01 | 2160 | 1.07 | 1.07 |
| | 2 | -0.16 | 0.01 | 2145 | 1.04 | 1.03 |
| | 3 | -0.13 | 0.01 | 2137 | 0.96 | 0.96 |
| 2 | 4 | -0.26 | 0.01 | 5909 | 0.99 | 0.98 |
| | 5 | 0.26 | 0.01 | 5851 | 1.37 | 1.46 |
| | 6 | 0.11 | 0.01 | 5829 | 0.96 | 0.96 |
| 3 | 7 | -0.13 | 0.01 | 5529 | 0.89 | 0.89 |
| | 8 | -0.40 | 0.01 | 5592 | 0.92 | 0.92 |
| | 9 | -0.16 | 0.01 | 5522 | 0.89 | 0.89 |
| 4 | 10 | -0.21 | 0.01 | 3025 | 0.96 | 0.95 |
| | 11 | -0.06 | 0.01 | 2957 | 0.92 | 0.91 |
| | 12 | -0.10 | 0.01 | 2908 | 1.01 | 1.01 |
| 5 | 13 | -0.18 | 0.01 | 4337 | 0.97 | 0.98 |
| | 14 | 0.21 | 0.01 | 4136 | 0.92 | 0.92 |
| | 15 | 0.10 | 0.01 | 4163 | 1.07 | 1.08 |
| 6 | 16 | 0.13 | 0.01 | 5665 | 1.02 | 1.02 |
| | 17 | -0.13 | 0.01 | 5702 | 1.02 | 1.02 |
| | 18 | -0.05 | 0.01 | 5637 | 1.08 | 1.08 |
| 7 | 19 | 0.20 | 0.02 | 701† | 1.08 | 1.08 |
| | 20 | 0.27 | 0.03 | 666† | 1.05 | 1.06 |
| | 21 | 0.31 | 0.03 | 599† | 1.10 | 1.11 |
| 8 | 22 | 0.40 | 0.03 | 369† | 1.01 | 1.02 |
| | 23 | 0.15 | 0.03 | 369† | 0.83 | 0.84 |
| | 24 | 0.12 | 0.03 | 343† | 0.92 | 0.91 |

† There are fewer than 10 counts in some of the higher categories for these items

* Mean square

While the item difficulty statistics provide information about the performance of the paper as a whole, threshold measures for each item illustrate the distribution of marks across the mark range. Threshold measures are normally presented in the form of probability plots which give the probability of a given mark being awarded, dependent upon the extent to which the latent trait is exhibited. To illustrate, consider once again Item 2 on Unit 1 of Specification X. Figure 5 shows the probability that each mark will be awarded. Close examination reveals that there is no point along the latent trait continuum at which a mark of six, four, two or one is the modal mark awarded. Put simply, whatever a candidate's ability these four marks are never the most likely to be awarded.

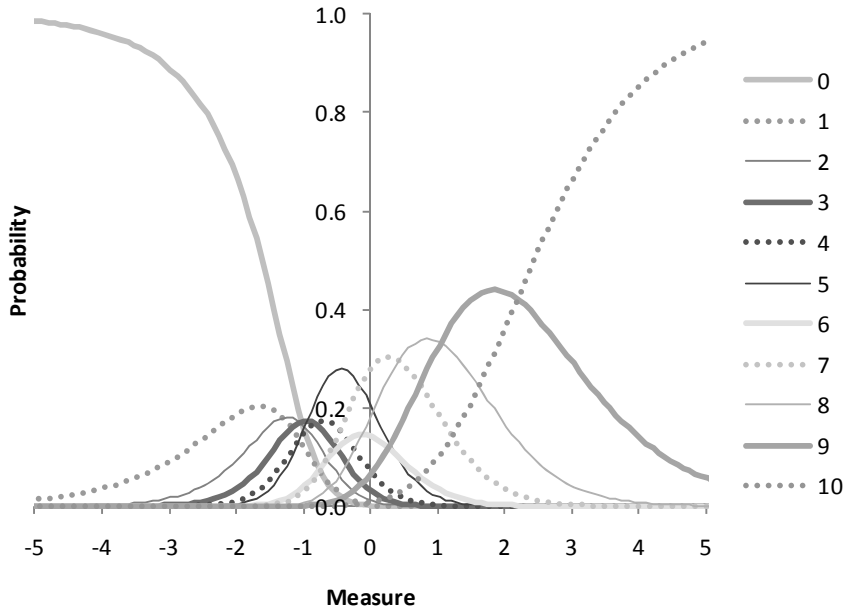


**Figure 5 Cumulative probability plot for Item 2 on Unit 1 of Specification X**

Although the relatively rare award of a mark of six was identified using the simple moving average model (Figure 3), the paucity of lower marks was not apparent using this model. While the additional information provided by a PCM probability plot might be useful in identifying shortcomings in a mark scheme, these plots are complex and difficult to interpret. Alternative presentations might provide the key for a wider audience. For example, the data might be formulated such that a bar chart shows the number of logits over which a particular mark is the modal award (Figure 6). Roughly the same number of logits for each mark might imply that each additional mark awarded had the same worth on the latent continuum[3]. Alternatively, the data might be formulated such that a line chart shows the range of the latent trait over which a given mark is modal (Figure 7). In Figure 7, a horizontal line denotes a modal mark and a vertical line denotes the transition between two, possibly non-adjacent, modal marks.

Both representations show that lower ability candidates are most likely to be awarded either zero or five marks. If the item can be assumed to measure a single latent trait, even though the marks of one, two, three and four are awarded, it appears that these marks are not very effective for differentiating between candidates.

[3] Because the latent trait is measured on a continuum, the minimum mark would be modal from -∞ to the first threshold and the maximum mark would be modal from the last threshold to +∞. Thus, these marks have been represented by arrows on Figure 6.

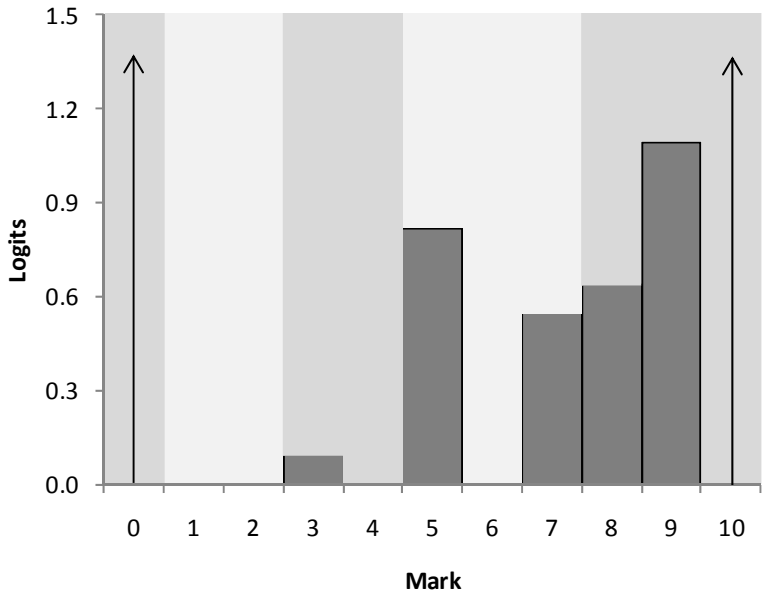


**Figure 6 Logits over which a given mark is modal for Item 2 on Unit 1 of Specification X**

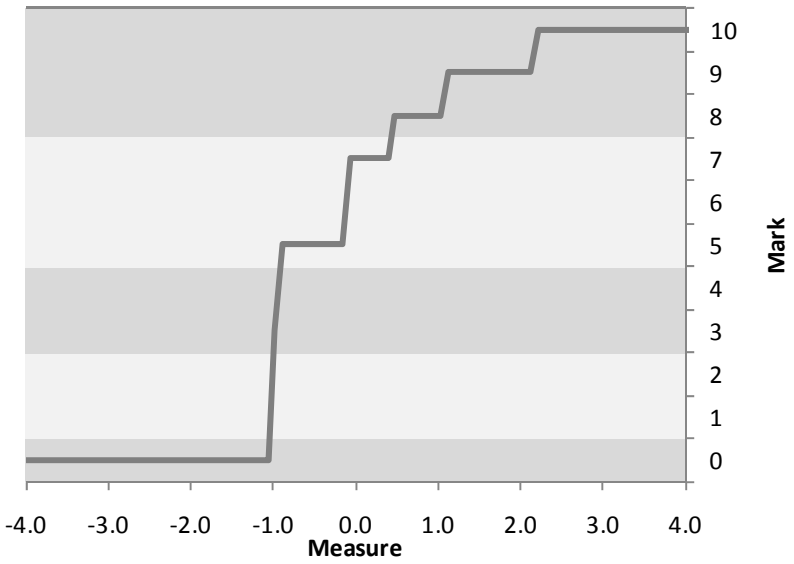


**Figure 7 Range of latent trait over which a given mark is modal for Item 2 on Unit 1 of Specification X**

## CONCLUSIONS

There has long been a focus on mark re-mark reliability when evaluating the effectiveness of mark schemes. While this has not been to the exclusion of recognising the importance of validity, techniques for assessing validity have not been well documented.

Among other things, a valid assessment must discriminate effectively between candidates and the mark scheme provides one instrument with which to achieve this aim. For each item on a paper, the issue of the value of a mark must underpin the design of the mark scheme. Ideally, 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. Thus, any item with underutilised marks has the potential to limit discrimination between candidates.

The identification of underutilised marks provides a key to resolving some of the potential inequities in an assessment. Techniques may range from the simple scrutiny of mark distributions through to the fitting of models, but whichever technique is used, there is an implicit reliance on some concept of the true mark distribution. Just as the true mark distribution will never be known, a perfectly reliable and valid assessment will never be achieved. Even with the ideal examination paper, it is unlikely that the mark's worth within each item and across all items would be equivalent. However, by attempting to identify weaknesses in a mark scheme, it might be possible to make refinements to allow for more effective discrimination between candidates and, ultimately, to build a set of rules for the design of mark schemes.

As a starting point, the simple statistics reported herein might be used as part of the current quality control mechanism whereby key information on the performance of every examination paper is fed back to the relevant committee (Stockford, Eason, & Taylor, 2010). Areas of concern, highlighted by the simple statistics, might give rise to the need for further investigation, with potential for the application of a PCM. In the future, a broader approach to designing assessments, using systems such as Outcome Space Control (Pollitt et al., 2008) or Evidence-Centred Design (Mislevy & Haertel, 2006), might reduce the prominence of post hoc evaluation and feedback. Nevertheless, the continued development of a suite of mechanisms for identifying failures within an assessment should further enhance the reliability and validity of the product offered.

## REFERENCES

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573.

Andrich, D. (1998). Thresholds, Steps and Rating Scale Conceptualization. *Rasch Measurement Transactions*, *12*(3), 648-649.

Andrich, D., de Jong, J. H. A. L., & Sheridan, B. E. (1997). Diagnostic opportunities with the Rasch Model for ordered response categories. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences*. New York: Waxmann.

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

Bramley, T. (2001). The Question Tariff Problem in GCSE Mathematics. *Evaluation & Research in Education*, *15*(2), 95-107.

Bramley, T. (2008). Mark scheme features associated with different levels of marker agreement. British Educational Research Association (BERA) annual conference. Heriot-Watt University, Edinburgh.

Linacre, J. M. (2002). Understanding Rasch Measurement: Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement*, *3*(1), 85-106.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Meadows, M., & Billington, L. (2006). *A review of the literature on marking reliability*. National Assessment Agency.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice*, *25*(4), 6-20.

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.

Opposs, D., & He, Q. (2011). *The reliability programme: Final report*. Office of Qualifications and Examinations Regulation.

Pollitt, A., Ahmed, A., Baird, J., Tognolini, J., & Davidson, M. (2008). *Improving the Quality of GCSE Assessment*. Qualifications and Curriculum Authority.

Stockford, I., Eason, S., & Taylor, R. (2010). Question Paper Functioning Reports. *AQA Internal Report*, *RPA_10_IS_MO_010*.

Anne Pinot de Moira
05 April 2011