Seeds of Doubt

Learning lessons from item level marking

Anne Pinot de Moira

Summary

This paper proposes a multilevel model which could be used to compare marking reliability between seed marked items and between seed marked units. It is hoped that the output from the model might promote the sharing of good practice and allow AQA to present a more unified approach to the design of mark schemes and items.

Keywords: marking reliability, seed data, multilevel models

Introduction

More ready access to item level information for units which are marked using the CMI+ system has allowed increased scrutiny of the reliability with which these items are marked. Data are available for each CMI+ unit, each seed within that unit and each occasion on which the seed was remarked; thus allowing comparison of the seed mark and remark to provide a measure of marking reliability.

Increases in reliability can be effected in many ways. It can be argued that mark reliability, in other words reliability of the final mark awarded to a candidate, can be increased by item level marking (see, for example, Wheadon & Pinot de Moira, 2012) or some form of multiple marking (see, for example, Meadows & Billington, 2005; Pilliner, 1968). In contrast, and here it is important to make a distinction, mark*ing* reliability can only be improved by examiner training and improvements to the assessment and mark scheme. This paper provides a mechanism by which to focus efforts on areas where changes to training and the assessment might improve marking reliability. Whilst technical in nature, the analysis in this paper is presented with the introduction of new operational practices in mind.

A model of marking reliability

The data

In 2012, there were 5,335,217 seed marked responses recorded during the marking window. The analysis of marking reliability presented here used a stratified sample of these responses. The sampling frame was restricted to include only items with a maximum mark tariff of two or more and to include a maximum of 500 responses per item (525,310 responses met these criteria). The final sample comprised 52,991 marks awarded to seed items and exemplified the work on 3,351 items from 235 units.

For each response, the absolute difference between the seed mark and the examiner mark was calculated. Because of the examiner hierarchy observed within the system and the fact that the seed mark is assigned by a senior examiner, the seed mark was assumed to represent the gold

AQA Education (AQA) is a registered charity (number 1073334) and a company limited by guarantee registered in England and Wales
(number 3644723). Our registered address is AQA, Devas Street, Manchester M15 6EX.

standard. Therefore, the difference from the seed mark was taken as a proxy for marking reliability: the smaller the difference, the more reliable the marking.

The multilevel model

To improve marking reliability, it is important to focus on the sources of error. It is relatively straightforward, by use of simple statistical measures, to identify examiners whose marking is at odds with the standard. However, errors are also likely to come from the assessment, the item or the mark scheme design as well as the examiners themselves. To gain a fuller picture of marking reliability requires modelling the interrelations between all these sources.

Thus, a three-level linear multilevel model was fitted to assess marking reliability for the differing items and units (Equation 1). Seed marked responses (*i*) were nested within items (*j*) and items were nested within units (*k*). The absolute difference between the seed mark and examiner mark was used as the dependent variable. The maximum mark for the item was included as an independent variable. The parameter estimates associated with the model are included in Appendix A.

$$\begin{aligned} Absolute\ Difference_{ijk} &= \beta_{0ijk}Cons + \beta_1 Item\ Maximum_{jk} \\ \beta_{0ijk} &= \beta_0 + v_{0k} + u_{0jk} + e_{0ijk} \\ v_{0k} &\sim N(0, \sigma^2_{v0}) \\ u_{0jk} &\sim N(0, \sigma^2_{u0}) \\ e_{0ijk} &\sim N(0, \sigma^2_{e0}) \end{aligned}$$

Equation 1

The model fit and limitations

Although this model represented a significant improvement on a null model it still explained very little variation in the data. R^2 was crudely estimated at about 18%.[1] Furthermore, the final model showed that the greatest proportion of unexplained variation was at the response level (61.1%).

Thus, attempts to improve marking reliability by identifying problem units and items may be limited by the vagaries of candidate responses or, in other words, the interaction between the examiner and the seed response. Nevertheless, it must surely be the responsibility of the awarding bodies to standardise this interaction. Standardisation can only be achieved by the careful design of items and mark schemes. That is not to say all items should be designed such that there is a closed set of responses as this would undermine the validity of the assessment. They should, however, avoid ambiguity and limit scope for multiple disparate, but ostensibly correct, responses.

Output

Analysis of the random effects allowed the ranking of items and units in terms of marking reliability. Figure 1 and Figure 2 show the simultaneous confidence intervals for the item and unit level residuals, respectively. A lower residual represents more reliable marking and a higher residual, less reliable marking. In Figure 1, the separate panels represent the different maximum mark tariffs. As might be expected, as the maximum mark for an item increases, so generally does the difference from the seed. For the lower tariff items, there is almost no difference in the marking reliability dependent upon the item being marked. All the confidence

[1] Estimated as $\left(1 - \left({}^{\text{Residual variance in the full model}}/{}_{\text{Residual variance in the null model}}\right)\right) x\ 100$

intervals overlap. For some of the higher tariff items, however, it is possible to discern differences between the most reliably and least reliably marked items. The data from which the graphs are derived might, therefore, be of use in identifying areas of good and poor practice in terms of item and mark scheme design, and of training.

Table 1 provides an extract from the table ranking the unit level residuals. It gives a clear illustration of why these data should be treated with caution. It should come as no surprise that units from subjects such as statistics and accounting appear towards the top as the most reliable, while more subjective specifications are at the bottom. Nevertheless, for the purposes of improving marking reliability, efforts might be best focussed towards anomalies or outliers.

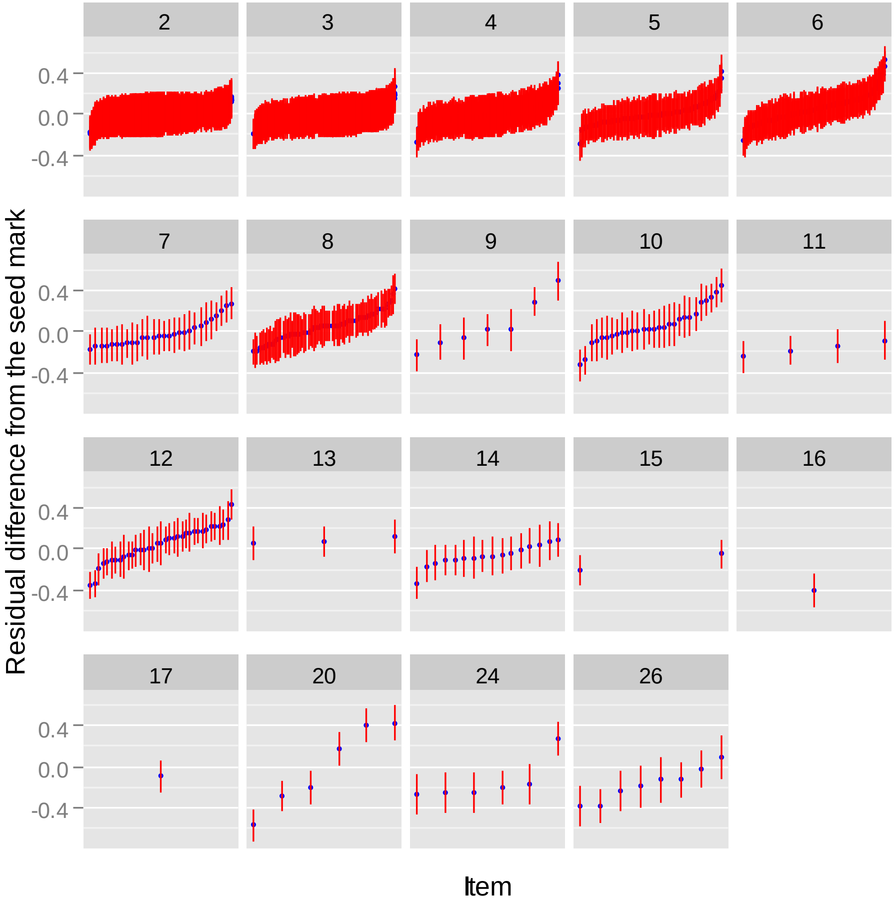

Figure 1 **Simultaneous confidence intervals for the item level residuals expressed as differences from the mean level of reliability (low value denotes high reliability)**

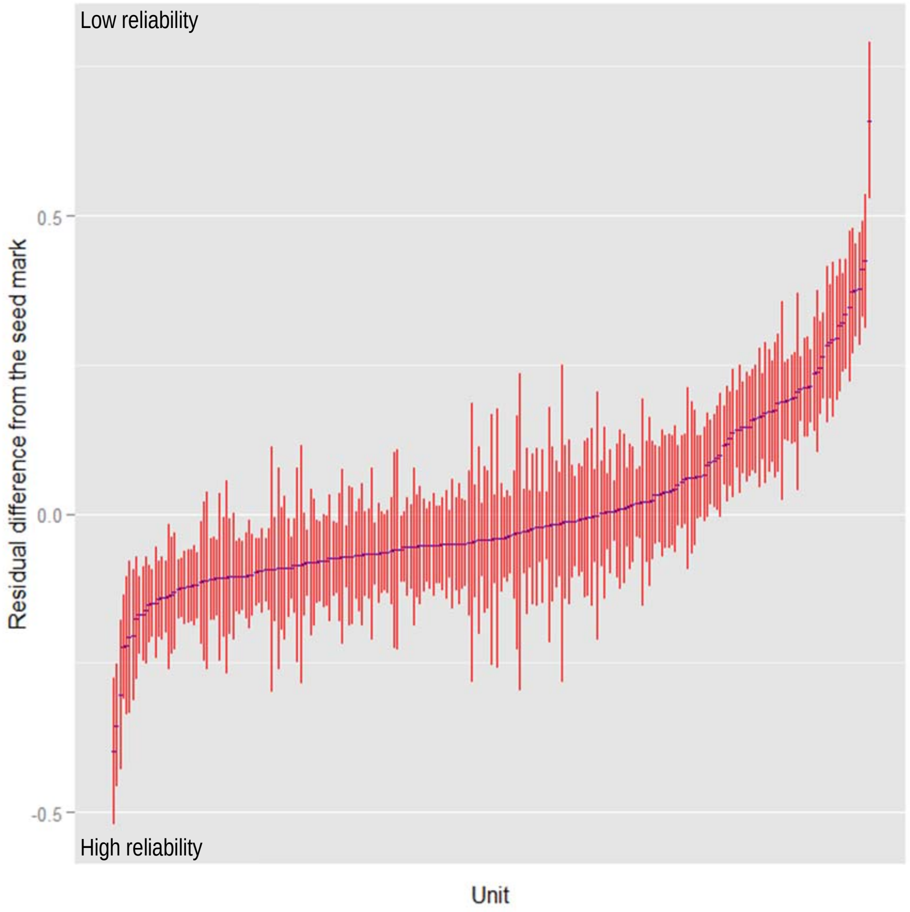

Figure 2 **Simultaneous confidence intervals for the unit level residuals expressed as differences from the mean level of reliability (low value denotes high reliability)**

Table 1 **Rank of unit level residuals**

Subject	Unit	Rank
Statistics	SS03	1
Accounting	ACCN2	2
Statistics	SS05	3
Statistics	SS06	4
:	:	:
:	:	:
D&T: Food Technology	FOOD1	232
Business Studies	BUSS2	233
Economics	ECON2-2	234
Citizenship	CIST1	235

1: relatively high reliability → 235: relatively low

While the intention of this paper is not to name and shame, rather to propose a method for identifying areas where marking reliability needs improvement, it is instructive to point to CIST1 where the combined information from Table 1 and Figure 2 suggest cause for further investigation. CIST1 has marking reliability which appears significantly poorer than any of the other units in the sample, as evidenced by the largely non overlapping confidence intervals.

A similar look at the item level residual ranks also provides interesting information. Take, for example, the items marked out of 24, all of which come from the religious studies specification

(Table 2). Figure 1 shows that this essay question was significantly more poorly marked for the Islam unit than for any of the other electronically marked units. Expressed in terms of marks, the mean difference from the seed was nearly half a mark greater than that for the comparable items. Could this be due to the nature of the question posed? Could it be a feature of mark scheme design? Could it be because there are difficulties with recruiting and retaining examiners with expertise in the area?

Table 2 Rank of item level residuals for items with a 24 mark tariff

Subject	Unit	Item	Rank
RS St Luke’s Gospel	405006	B5_6	1
RS Judaism	405010	B5_6	2
RS Judaism: Ethics	405011	B5_6	3
RS Islam: Ethics	405009	B5_6	4
RS Buddhism	405012	B5_6	5
RS Islam	405008	B5_6	6

(1: relatively high reliability → 6: relatively low reliability)

Conclusions and recommendations

Identifying the *causes* of differences in marking reliability between units and between items was beyond the scope of this paper. What this paper aimed to provide was a mechanism by which to spot areas of concern. A trained eye, for example, might be able to return to the mark schemes for CIST1 and, by comparison with similar subjects with proven good practice, suggest changes which could increase marking reliability. At best, any increases in marking reliability would be incremental and possibly unquantifiable but, nevertheless, the act of comparison across items, units and specifications should result in a more unified offering from AQA.

In the future, marking reliability information in this form might be added to the armoury of post results statistics which are provided to qualification developers. Ideally, the information would be used to encourage the sharing of good practice, rather than the identification of poor practice. It might also help in the targeting of resources for examiner recruitment and training.

Although the model might have been improved by including more explanatory variables, the fact that most of the unexplained variance was at the response level suggests that such an exercise would be of little additional value in identifying problems with marking reliability. However, as the analysis presented was based on a relatively small sample of seed marked items, it would be beneficial to increase this sample size were the resultant data to be used operationally. Furthermore, perhaps greater gains might be made from the analysis of sample double marked items since it is these items which have the higher mark tariff and therefore greater susceptibility to marking reliability problems.

References

Meadows, M., & Billington, L. (2005). *A review of the literature on marking reliability.* Report produced for the National Assessment Agency.

Pilliner, A. E. G. (1968). Examinations. In H. J. Butcher (Ed.), *Educational research in Britain*. London: University of London Press.

Wheadon, C. & Pinot de Moira, A. (2012). *Gains in marking reliability from item-level marking: Is the sum of the parts better than the whole?* CERP_TR_CBW_01102012. Manchester, UK: AQA Centre for Education Research and Policy.

Appendix A Parameter estimates for the model described in Equation 1

Effects	Parameter	β	se(β)	p
Fixed	Constant	-0.050	0.011	0.000
	Item Maximum	0.069	0.001	0.000
Random	Unit	0.115	0.011	0.000
	Item	0.151	0.011	0.000
	Response	0.417	0.000	0.000

Explained variance R^2 = 0.180