Quality control of marking: Some models and simulations

John F. Bell, Tom Bramley, Mark J. A. Claessen and Nicholas Raikes*

Cambridge Assessment, 1 Hills Road, Cambridge CB1 2EU, United Kingdom

*Corresponding author. Email: Raikes.N@CambridgeAssessment.org.uk

Abstract

As markers trade their pens for computers, new opportunities for monitoring and controlling marking quality are created. Item-level data may be collected and analysed throughout marking. The results can be used to alert marking supervisors to possible quality issues earlier than is currently possible, enabling investigations and interventions to be made in a more timely and efficient way. Such a quality control system requires a mathematical model that is robust enough to provide useful information with initially relatively sparse data, yet simple enough to be easily understood, easily implemented in software and computationally efficient – this last is important given the very large numbers of candidates assessed by Cambridge Assessment and the need for rapid analysis during marking. In the present paper we describe the models we have considered and give the results of an investigation into their utility using simulated data.

Introduction

New technologies are facilitating new ways of working with examination scripts. Paper scripts can be scanned and the images transmitted via a secure Internet connection to markers working on a computer at home. Once this move from paper to digital scripts has been made, marking procedures with the following features can be more easily implemented:

- Random allocation: each marker marks a random sample of candidates;
- Item level marking: Scripts are split by item – or by groups of related items – for independent marking by different markers;
- Near-live analysis of item-level marks: item marks can be automatically collected and collated centrally for analysis as marking proceeds.

Features such as these open the possibility of analysing item marks during marking and thereby alerting supervisors to possibly aberrant marking earlier than would otherwise be possible.

In the present paper we consider the following types of aberrancy, although the models and methods we discuss could be applied to other forms of marker aberrancy:

- Overall severity: the marker is more severe or less severe (i.e. lenient) than he or she should be over all items;
- Item-specific severity / leniency: the marker's severity varies by item, e.g. the marker might be lenient on one item but severe on another;
- Erraticism: the marker marks erratically, perhaps due to carelessness; this is a random error that might apply to a marker's marking of one or more particular items or across all items.

We consider two numerical models, a three facet, partial credit Rasch model (See Linacre, 1989, for technical details), and a simpler model based on generalizability theory (see Shavelson & Webb, 1991) that we refer to for convenience as our 'means model'.

The reader may wonder why we developed a simple model if a Rasch model could be used. Our reasons relate to the environment in which we propose the model be used: near-live, repeated analyses of many datasets that are initially sparse but can become very large indeed. In these circumstances, the drawbacks of a partial credit, multi-facet Rasch model include:

- The amount of computationally intensive, iterative processing needed;
- The difficulty and cost of implementing such a relatively complex model in a reliable 'back office' examination processing system suitable for routine use in a high volume, high stakes, high pressure environment;
- The lack of a body of evidence on which to rest assumptions concerning the validity of the Rasch model when applied to many of the question papers used by Cambridge Assessment, which typically intersperse items of many different types and numbers of marks, and where reverse thresholds are often encountered;
- The difficulty of explaining the model to stakeholders with little or no technical knowledge;
- The fact that the estimation of Rasch parameters is an iterative process, and different convergence limits might need to be set for different data sets. This could affect the residuals, which in turn affect whether a particular piece of marking is flagged as potentially aberrant.

We therefore decided to develop a much simpler model, and compare its performance with that of a multi-facet, partial credit Rasch model, using a range of simulated data.

The Data Simulator

Two overriding considerations led to our use of simulated data: the ability to produce large volumes of data at will, and the ability to control the types and degree of aberrance and thus facilitate systematic investigation of the models to an extent not possible with real data.

The data are generated from a 3-facet model:

$$\ln\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = B_n - D_i - S_{ij} - F_{ik}$$

where:

P_{nijk} is the probability that marker *j* gives candidate *n* a mark of *k* on item *i*

P_{nijk-1} is the probability that marker *j* gives candidate *n* a mark of *k-1* on item *i*

B_n is the 'ability' of candidate *n*

D_i is the 'difficulty' of item *i*

S_{ij} is the 'severity' of marker *j* on item *i*

F_{ik} is the threshold marking the transition from score category *k*-1 to score category *k* on item *i*.

B, D, S and F can all be considered as locations on a single latent trait or dimension. This model is therefore a unidimensional model. The fact that the term for the marker facet (S_{ij}) contains both the item and the marker allows the simulation of markers whose severity varies depending on the item. This interaction means that the above model is not a Rasch model, which would require the contribution of each facet on the right-hand side of the equation to be additive (i.e. without any interaction terms). However, the generator can still simulate data according to the Rasch model by fixing the severity of each marker across all items which they are eligible to mark.

Rather than generating a full set of thresholds (F_k) for each item, we use the solution of Andrich & Luo (2003) and generate just the first four principal components (mean,

standard deviation, skew and kurtosis) of the distribution of thresholds, and derive the individual thresholds from these components. A full mathematical treatment of this approach is beyond the scope of this paper, but the reader can consult Andrich & Luo (2003) for details.

The configuration options governing the generation of candidate ability parameters, marker severity parameters and item difficulties are as follows.

Candidates: This is the simplest part of the generator. We specify the number of candidates N, and the mean M and standard deviation S of their ability distribution (in logits – the log-odds unit of the Rasch model).

A total of N candidate ability parameters (B) are then generated from a normal distribution with mean M and standard deviation S.

I.e. B ~ Normal (M,S)

Markers: We specify the severity in logits of each marker on each item. A value of zero means neither severe nor lenient, positive values indicate increasing severity and negative values indicate increasing leniency (a missing value indicates that we do not wish to generate data for that marker on that item, i.e. the marker 'did not mark' that item).

We also specify the erraticism in marks of each marker on each item. This number specifies the standard deviation of a normal distribution with mean zero from which an error value for that marker on that item will be drawn at random. This value is then rounded to whole marks. If the marker is not 'erratic' on that item then the value will be zero.

Items: This is the most complex part of the data generator. The problem is to generate items which vary realistically in both overall difficulty and in the distribution of marks across the item, given that the generator needs to be capable of producing output for items with widely varying numbers of marks.

For each item on the test, the generator takes as its input:

- A unique item ID;
- The serial position of the item in the test;
- The maximum number of marks available, and
- The 'item group' it is in. This last feature allows for several items to be grouped together as a block which becomes the smallest unit for allocation to markers.

The method we used for generating the item parameters made use of a database containing more than 900 items from many subjects, their principal components compiled from past Rasch analyses of real data. Given an item worth *m* marks, a set of component parameters was obtained by sampling once at random from the subset of items in the database worth *m* marks. This ensured that the four parameters were 'realistic' in that they have been obtained in the past.

In fact, analysis of the database suggests that the first component (item difficulty) is relatively independent of the other components so in the data generator this component is obtained by random sampling from the normal

distribution[1], and the other three components are obtained as a set by sampling from the database[2].

Having obtained our sets of simulated candidate, item and marker parameters, the final stage is to simulate the marks. First each candidate is randomly allocated a marker on each item, subject to any constraints specified for item grouping and marker eligibility. Next the simulated marks are generated for each candidate on each item using the equation for the 3-facet model as follows. First of all, the set of cumulative probabilities for each possible mark on an item for a given candidate's ability is generated (i.e. for each possible mark on the item [0,1, …, max], the probability of the candidate getting that mark or lower is calculated). Next a number between 0 and 1 is drawn at random from a flat distribution and this random number is compared with the cumulative probabilities. The simulated mark x for the given candidate on the given item is the mark where the random number lies between the cumulative probability for that mark and the cumulative probability for the mark above. The procedure is repeated for all candidates over all items.

Here is an example for a four mark item:

Cumulative probabilities for cand. n on item i marked by marker j					Random number	Simulated mark (before erraticism applied)
P(0)	P(1)	P(2)	P(3)	P(4)	r	x
0.1	0.3	0.4	0.9	1.0	0.4256782	2

Finally, marker erraticism is applied. A number is drawn at random from a normal distribution with mean 0 and the specified standard deviation. This number is added to the simulated mark and the total rounded to the nearest whole mark or, if this would result in a mark beyond the permitted range, to zero or full marks as appropriate.

The means model

Our simple model is not a rigorous statistical model. Its intended purpose is to flag markers whose marking patterns deviate in some way from the majority of markers, suggesting – but not proving – a degree of aberrancy on the part of the marker. In this way senior examiners' checks on marking quality might be prioritised so that they first review the marking most likely to be aberrant, thereby cutting the time taken to detect, diagnose and remedy aberrant marking.

This is still a work in progress and the model has not been finalised. We use generalizability theory to partition candidates' marks into a series of effects -- see Shavelson & Webb, 1991, for technical details.

The examination we used in our investigations

We chose a Leisure and Tourism examination because this examination contained a wide range of types of item, and because it is likely to be marked on screen in the near future and so real data will be available against which our simulated data may be compared.

---

[1] The available evidence suggests that difficulties are noticeably more spread out for 1 and 2-mark items so the data generator takes this into account.

[2] The database does not yet contain a significant number of items worth more than 7 marks. Therefore all four principal components of the distribution of thresholds are generated randomly for items worth more than 7 marks, though the range of each component is based on the range found in the database.

The examination consists of four questions, each of which contains four parts, a, b, c and d, worth 4, 6, 6 and 9 marks respectively.

The part (a) items are essentially objective, for example asking candidates to select four pieces of information matching a given criterion from a larger list of given information. Markers do not need domain-specific knowledge to mark these items.

Part (b) items are more open ended, for example asking candidates to explain three things and giving, for each one, the first mark for a reason and the second for an explanation. Markers need some domain-specific knowledge to mark these items.

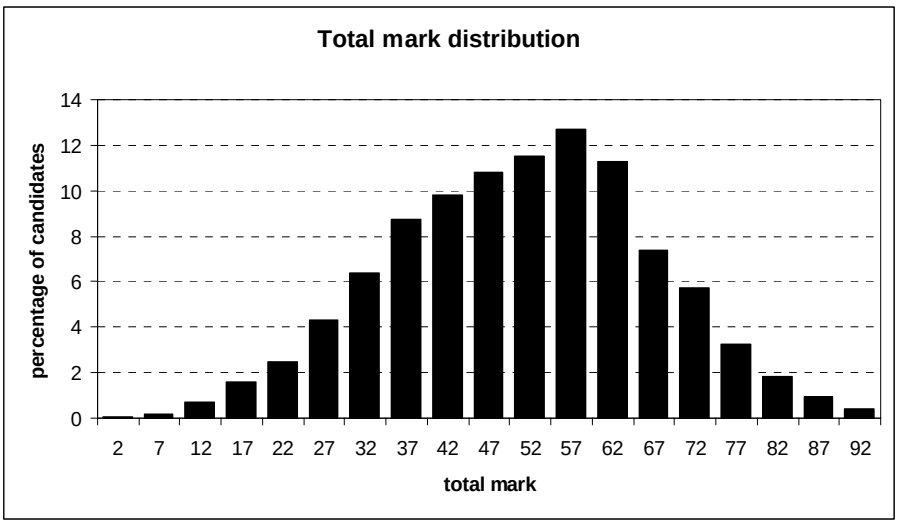
Part (c) and (d) items required candidates to write more extended answers, which are marked holistically against 'levels of response' criteria, the mark scheme containing a brief description of each level of response. Again, markers need domain-specific knowledge for these items.

How realistic are the simulated data?

For our first, baseline simulation, we simulated Leisure and Tourism data for 3200 candidates. Their mean ability was set to zero logits, and the standard deviation of their abilities was set to 0.69 logits. The simulation contained no marker severity or erraticism, only random error. Apart from the (c) and (d) items within each question, which were kept together in an item-group, scripts were simulated to have been split by item and the answers distributed to markers at random. That is, each of the four questions was split into three item-groups, (a), (b) and (c,d), making twelve item groups in total. Answers to these item-groups were distributed to markers at random.

The resulting distribution of simulated total marks is shown in Figure 1. The mean mark was 51.0 and the standard deviation was 15.7.

Figure 1: Distribution of total marks, simulation 1



Classical item statistics calculated from the simulated data are given in Table 1. We do not currently collect real item level data for Leisure and Tourism, so the simulated data cannot be compared with real data. However, experience from other examinations suggests the items have a fairly realistic range of facility values, and the later questions are more difficult than the earlier questions, as is often the intention. The items are all discriminating well, and Cronbach's alpha (internal consistency reliability) is high.

Table 1: Item analysis for simulation 1

Item	Item label	Max mark	Item mean	Item SD	Facility	Discrimination
1	Q1a	4	3.34	0.90	0.83	0.45
2	Q1b	6	4.46	1.37	0.74	0.56
3	Q1c	6	3.81	1.41	0.64	0.57
4	Q1d	9	4.28	2.69	0.48	0.71
5	Q2a	4	2.20	1.19	0.55	0.53
6	Q2b	6	3.46	1.50	0.58	0.60
7	Q2c	6	2.31	1.46	0.38	0.60
8	Q2d	9	7.05	1.72	0.78	0.55
9	Q3a	4	2.28	1.18	0.57	0.54
10	Q3b	6	2.77	1.51	0.46	0.61
11	Q3c	6	2.80	1.93	0.47	0.65
12	Q3d	9	4.60	1.40	0.51	0.58
13	Q4a	4	1.31	1.01	0.33	0.49
14	Q4b	6	1.96	1.53	0.33	0.61
15	Q4c	6	1.26	1.57	0.21	0.56
16	Q4d	9	3.13	1.66	0.35	0.59
Chronbach's alpha= 0.90						

Detecting overall marker severity / leniency

We simulated the effects of adding overall marker severity to the baseline. Sixteen markers were simulated, all of whom marked all items. Each marker was simulated to have the same severity on all items, and the markers ranged in severity from -0.40 logits to 0.40 logits in intervals of 0.05 logits. Each marker was also simulated to have an erraticism of 0.2 logits on all items[3]

Overall marker leniencies were estimated using the means model – we have referred to the effect as 'leniency' because higher values mean higher marks. The overall marker severities were also estimated using the partial credit, three facet Rasch model. The results are shown in Figures 2 and 3 respectively. Each cross represents a marker, and the dotted line represents the situation where the estimated severities are perfectly reproduced, i.e. ignoring the added component of random error in the simulated data. Note that the means model estimates leniency in marks, a non-linear scale, whereas the Rasch model estimates severity on a linear logit scale. The Rasch model has done a good job in recovering the simulated severities, with all markers in the correct rank order. The means model has done almost as well, however, with only a few small 'mistakes' in rank order near the middle of the range – for the purposes of flagging potentially aberrant marking for investigation this is of negligible importance.

Figure 2: Means model - estimated leniency as a function of simulated severity

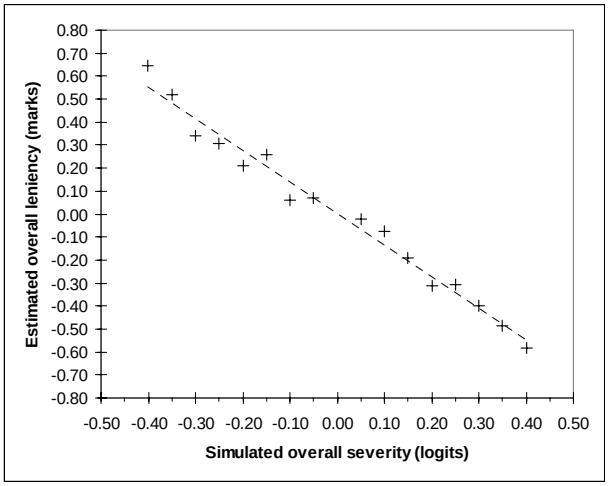
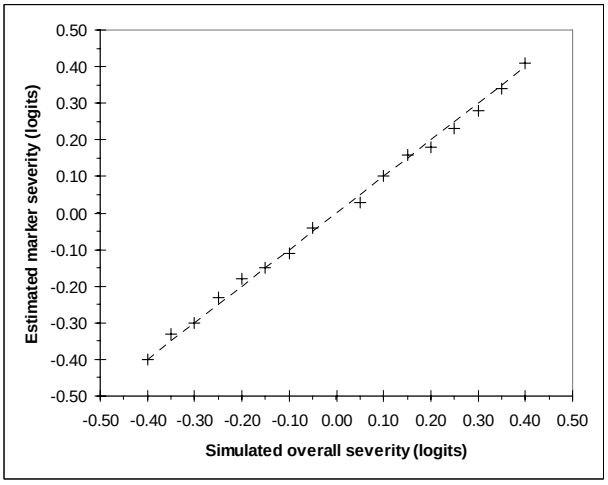


Figure 3: Rasch model: estimated severity as a function of simulated severity



[3] For this simulation, erraticism was added to a marker's severity in logits before a mark was simulated for an answer. The figure of 0.2 specified the standard deviation of a normal distribution from which an error value for that marker for that answer was drawn at random and added to the marker's severity. This approach to adding marker erraticism was replaced in later simulations by the method outlined on page 4 under 'marker' because the erraticism achieved depended on item difficulty.

Detecting item-specific severity

Sometimes a marker may consistently mark a particular item or items more severely or leniently than other items. This can be detected as marker-item bias. Observed biases may be the result of several causes. For example, if a marker marks a mixture of high judgement and low judgement items, any severity or leniency might only be apparent on the high judgement items. Alternatively, if the marker misunderstands the mark scheme for a low judgement item, he or she may consistently give too many or too few marks to every answer that fits his or her misunderstanding. Both these sources of bias can be simulated by considering markers to have item-specific severities. Another, more subtle source of marker-item bias occurs only for difficult or easy items, when an erratic marker might appear biased since his or her errors cannot result in a mark more than an item's maximum mark or less than zero.

We investigated the effects of adding some item-specific severities to our simulation. We divided our markers into two groups, following a realistic divide: the essentially objective part (a) items were marked by six General Markers; the other items, which required markers to have domain specific knowledge, were marked by twelve Expert Markers. All the General Markers' severities were simulated to be zero for all their items. Each Expert Marker was simulated to be severe or lenient by 0.5 logits on one item. All markers were simulated to have an erraticism of 0.1 marks on all items.

Marker-item biases were estimated from the means model, and the partial credit, three facet Rasch model. The results are shown for Expert Markers only in Figures 4 and 5 respectively. A triangle denotes a marker who was simulated to be severe by 0.5 logits on an item, a circle denotes a marker simulated to be 0.5 logits lenient on an item, and a cross denotes markers whose simulated item-specific severities were zero. It can be seen that both the means model and the Rasch model clearly distinguished the aberrant marker in each case.

Figure 4: Means model - marker-item bias

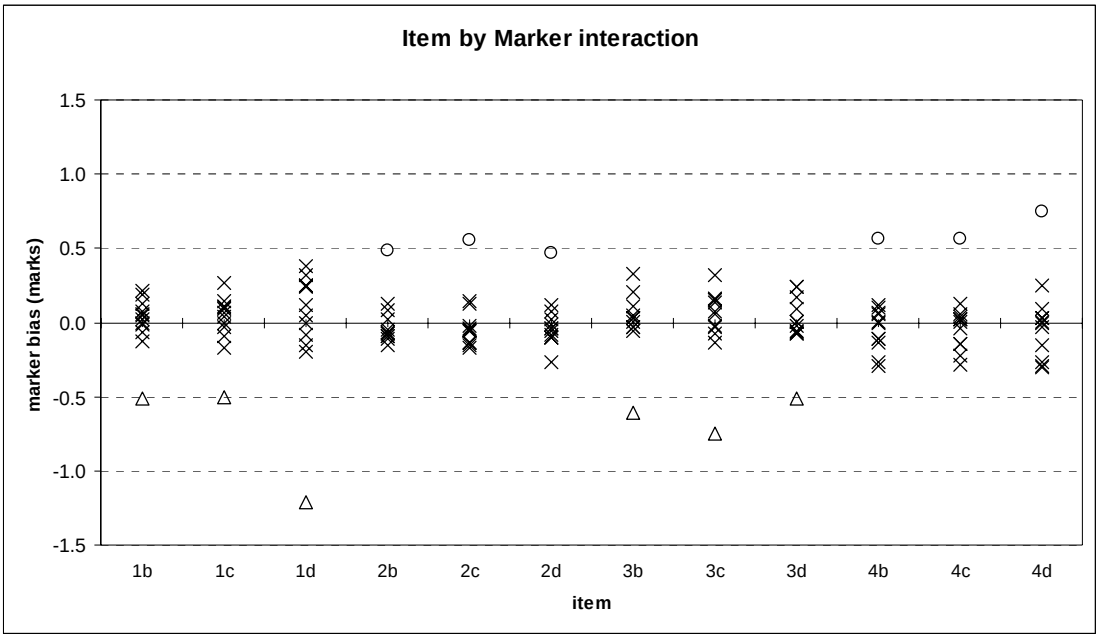
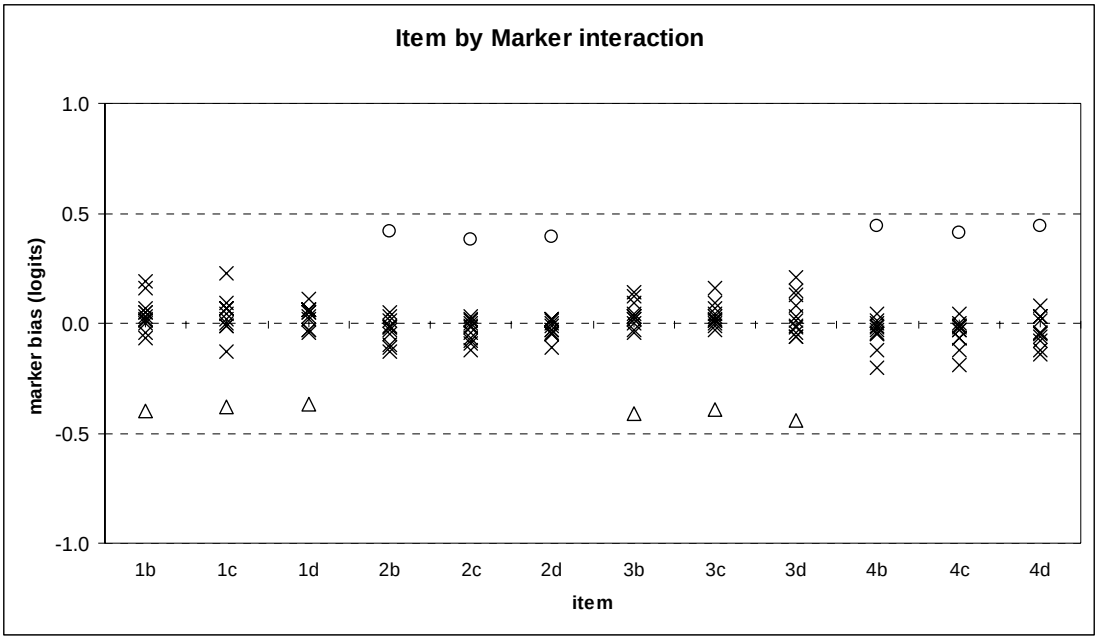


Figure 5: Rasch model - marker-item bias



Detecting erraticism

In a simulated data set for 3200 candidates with no severe or lenient markers, one marker was made erratic on each item (parameter = 0.5; see page 4 under 'markers'). As before, the part (a) items were marked by six General Markers, and the other items were marked by twelve (different) Expert Markers. Each question was split into three item-groups as before and the answers to these item-groups were distributed at random to eligible markers.

For our experiments in detecting erraticism, we took the simplest possible approach, and just compared the markers' variances on each item. The hypothesis was that erratic markers would have larger variances. Figures 6 to 9 summarise the results – the deviation of each marker from the mean variance for the item in question is plotted. Triangles denote the marker with the simulated erraticism. It can be seen that for the (a) and (b) items, the erratic marker had the largest variance in each case – though not always by much. For the (c) and (d) items, there is no clear picture. From this evidence it would seem that relative variances might be useful for flagging potentially erratic markers of low tariff items, but that the situation is more complex for higher tariff items.

Figure 6: Relative variances - general markers

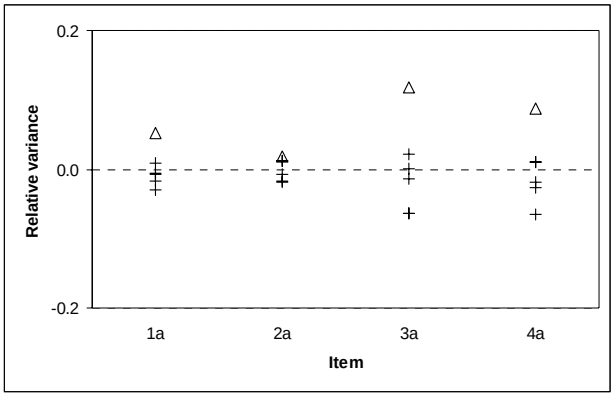


Figure 7: Relative variances - expert markers

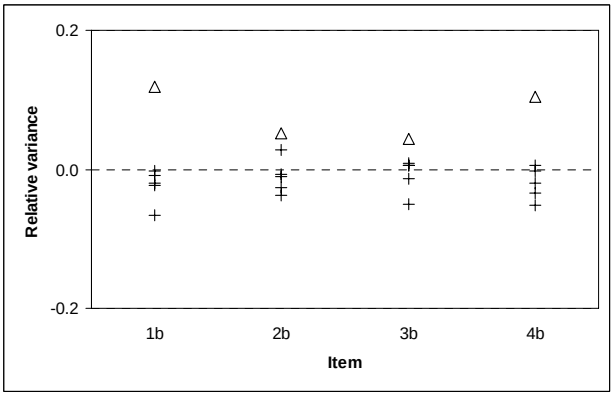


Figure 8: Relative variances - expert markers

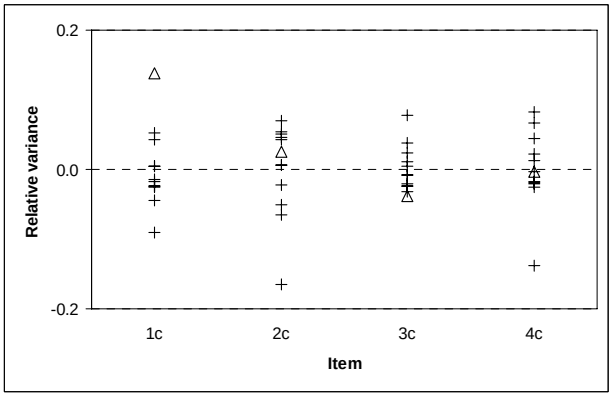
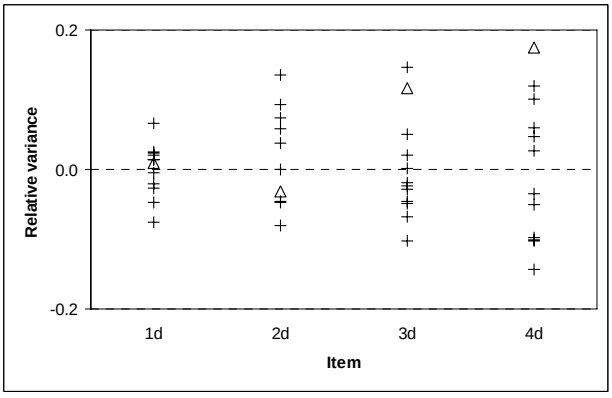


Figure 9: Relative variances - expert markers



Conclusion

Despite its computational simplicity, the means model has in these simulations proven itself capable of identifying severe and lenient markers, both ones that were severe or lenient across the board, and ones that were severe or lenient on particular items. When severities and leniencies were spread across a wide range, the means model was able to accurately rank order markers in terms of their severity and leniency, especially toward the extremes of the scales, where it matters most. The more complex and computationally demanding partial credit, multi-facet Rasch model that we used as a comparator offered little practical advantage in terms of the accuracy of the estimates it produced, especially when the purpose of the analysis is to prioritise marking for checking by a senior examiner.

On this basis, the means model seems very promising, and we are doing further work to validate the model with real data.

The simplest possible approach to detecting marker erraticism, namely distributing answers at random to eligible markers and comparing markers' variances, had mixed results. It would appear to be of some use for low tariff items, but of little use for higher tariff ones. A more fruitful approach might be to compare the variances of the residuals from the means model, and we will do further work to investigate this.

References

Andrich, D. & Luo, G. (2003). Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. *Journal of Applied Measurement, 4*(3), 205-221.

Linacre, J.M. (1989). *Many-facet Rasch measurement.* Chicago, MESA Press.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability Theory: A Primer*. Newbury Park, NJ, Sage Publications.