social, personal and affective responses. Aspects of markers' social histories as examiners and teachers were evident in the comparisons that they made and perhaps more implicitly in their evaluations. The overlap of these findings with aspects of various previous findings (e.g. the marking strategies identified by Greatorex and Suto, 2006) helps to validate both current and previous research, thus aiding the continued development of an improved understanding of the judgement processes involved in marking.

References

Baird, J., Cresswell, M. & Newton, P. (2000). Would the real gold standard please step forward? *Research Papers in Education*, **15**, 2, 213–229.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, **7**, 31–51.

Greatorex, J. (2000). *Is the glass half full or half empty? What examiners really think of candidates' achievement*. A paper presented at the British Educational Research Association Annual Conference, Cardiff, available at: http://www.leeds.ac.uk/educol/documents/00001537.doc (accessed 9 January 2007).

Greatorex, J. & Suto, W. M. I. (2006). *An empirical exploration of human judgement in the marking of school examinations*. A paper presented at the International Association for Educational Assessment Conference, Singapore, 2006.

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Cresskill, NJ: Hampton Press.

Laming, D. (2004). *Human judgment: the eye of the beholder*. London: Thomson.

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

Milanovic, M., Saville, N. & Shuhong, S. (1996). A study of the decision making behaviour of composition-markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment*. Cambridge: Cambridge University Press.

O'Hara, K. & Sellen, A. (1997). A comparison of reading paper and online documents. In S. Pemberton (Ed.), *Proceedings of the conference on human factors in computing systems*. 335–342. New York: Association for Computing Machinery.

Pula, J. J. & Huot, B. (1993). A model of background influences on holistic raters. In M. M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Cresskill, NJ: Hampton Press.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, **13**, 2, 191–209.

Sanderson, P. J. (2001). *Language and differentiation in Examining at A Level. PhD Thesis*. Unpublished doctoral dissertation, University of Leeds, Leeds.

Spear, M. (1997). The influence of contrast effects upon teachers' marks. *Educational Research*, **39**, 2, 229–233.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, **185**, 1124–1131.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts*. Norwood, N.J.: Ablex Publishing Corporation.

Wolf, A. (1995). *Competence based assessment*. Buckingham: Open University Press.

Quality control of examination marking

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

Abstract

As markers trade their pens for computers, new opportunities for monitoring and controlling marking quality are created. Item-level marks 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 article we describe the models we have considered and give the results of an investigation into their utility using simulated data.

This article is based on a paper presented at the 32nd Annual Conference of the International Association for Educational Assessment (IAEA), held in Singapore in May 2006 (Bell, Bramley, Claessen and Raikes, 2006).

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 identifying patterns that might indicate aberrant marking. Our aim is to speed up the detection of aberrant marking by directing marking supervisors' attention to the marking most likely to be

aberrant. In this way it will be possible to nip problems in the bud and reduce to a minimum the amount of marking that must be reviewed or re-done.

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

- Overall severity/leniency: the marker is consistently severe or lenient on all items.
- Item-specific severity/leniency: the marker's severity varies by item, for example, the marker might be lenient on one item but severe on another, or severe on one item but neutral on all others, etc.

It might be supposed that both of these types of aberrance could be satisfactorily remedied by applying overall or item-specific scaling factors to a marker's marks after all marking has been completed. If scaling is to be used, the results of the analysis would be used to help determine the appropriate scaling factors, rather than as a basis for intervention during marking. In many situations, however, scaling may be hard to justify, as in the case, for example, where a marker of factual items is severe because he or she is failing to reward some correct alternative answers. In these circumstances scaling is inappropriate and interventions must be made during marking if we are to avoid having to re-mark a considerable number of answers.

We consider two numerical models in the present paper: a three facet, partial credit Rasch model (see Linacre, 1989, for technical details); and a simpler model based on generalizability theory (see Shavelson and 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 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 (Andrich, de Jong and Sheridan, 1997) 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.

Why use simulated data?

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 basic process of evaluating a model using simulated data is:

1. Simulate the effects of particular kinds and degrees of marker aberrancy on a set of marks.
2. Analyse these simulated marks using the model being evaluated.
3. See whether the model correctly flags the simulated aberrancies.

Our simulator generates marks given the following configurable parameters:

- The number of candidates.
- The mean and standard deviation of the candidates' ability distribution in logits, the log-odds unit of the Rasch model.
- The severity in logits of each marker on each item. A value of 0 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).
- The 'erraticism' in marks of each marker on each item. Individual markers may vary in their consistency and this may also vary by item. The 'erraticism' parameter 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 for each answer marked. This value is then rounded to whole marks and added to the original (i.e. without erraticism) simulated mark.
- The number of marks m available for each item.
- Rasch item parameters for each item.

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 based our investigations on a question paper from GCSE Leisure and Tourism. We chose this question paper because it contained a wide range of types of item, and because some data from real marking was likely to become available against which the simulated data could be compared.

The question paper 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.

Our first, baseline simulation

For our first, baseline simulation, we simulated Leisure and Tourism data for 3,200 candidates. Their mean ability was set to 0 logits, and the standard deviation of their abilities was set to 0.69 logits. The baseline simulation contained no marker severity or erraticism, only random error. All markers were simulated to mark all items. Scripts were simulated to be split by item for marking, although within each question, items (c) and (d) were not split up. Answers were simulated to be distributed at random to markers.

Detecting overall marker severity/leniency

We simulated the effects of adding overall marker severity to the baseline simulation. Sixteen markers were simulated, all of whom marked all items. Each marker was simulated to be consistently severe or lenient across 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.

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 1 and 2 respectively. Each cross represents a marker, and the dotted line represents the situation where the estimated severities are perfectly reproduced. 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 – these small errors around 0 are of negligible importance, irrespective of whether the means model is to be used for the purposes of prioritising potentially aberrant marking for investigation, or for determining scaling factors.

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 items requiring different degrees of judgement to mark, 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 simulated data. We divided our markers into two groups, following a realistic divide: the essentially objective part (a) items were marked by one group of six markers (called the 'General Markers' hereafter); the other items, which required markers to have domain specific knowledge, were marked by a different group of twelve markers (referred to as 'Expert Markers'). All the General Markers' severities were simulated to be 0 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 from the partial credit, three facet Rasch model. The results are shown for

Figure 1 : Means model – estimated leniency as a function of simulated severity

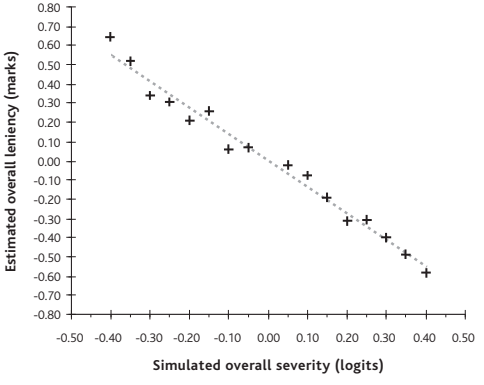
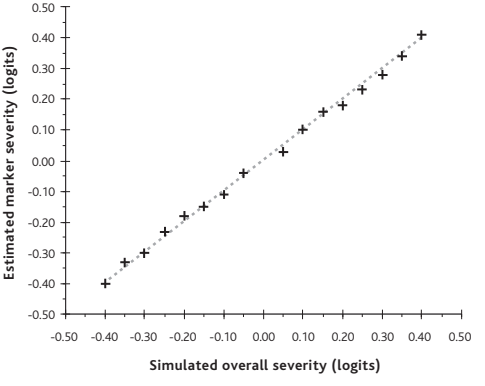


Figure 2 : Rasch model – estimated severity as a function of simulated severity



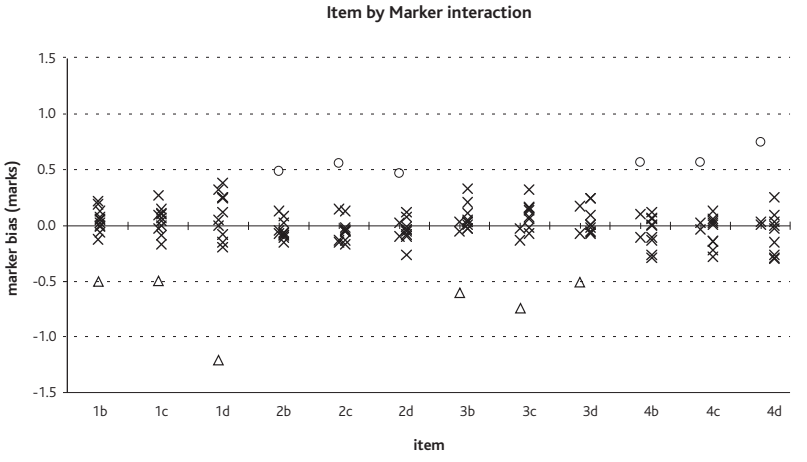


Figure 3: Means model – marker-item bias

Key:

△ = marker simulated to be severe by 0.5 logits on item

○ = marker simulated to be lenient by 0.5 logits on item

× = marker whose simulated item-specific severities were zero

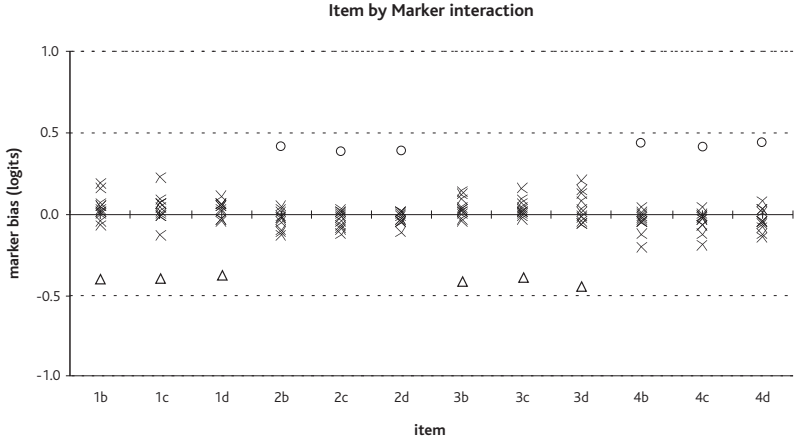


Figure 4: Rasch model – marker-item bias

Key:

△ = marker simulated to be severe by 0.5 logits on item

○ = marker simulated to be lenient by 0.5 logits on item

× = marker whose simulated item-specific severities were zero

Expert Markers only in Figures 3 and 4 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.

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.

References

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.), *Application of Latent Trait and Latent Class Models in the Social Sciences*, 59–70. Available at http://tinyurl.com/2eopcr

Bell, J. F., Bramley, T., Claessen, M. J. A. & Raikes, N. (2006). *Quality control of marking: some models and simulations*. Paper presented at the 32nd annual conference of the International Association for Educational Assessment, 21–26 May 2006, Singapore.

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.