Can Australians Mark KS3 Mathematics Exams? A Study in Cultural Differences

Paul Ayres

University of New South Wales, Sydney, Australia. (p.ayres@unsw.edu.au)

Highly experienced Australian teachers (N = 38) marked a sample of the 2006 KS3 mathematics exams, following similar training to their counterparts in the United Kingdom. Results indicated that were able to mark at a very high standard, but experienced a number of difficulties in doing so. Marking diaries revealed that a number of cultural differences existed concerning quality control, the number of questions marked, the different emphases attached to calculation accuracy, mathematical processes and conceptual understanding.

Introduction

High stakes examinations have for a long time been controversial. A major criticism of such examinations is that they lead to restricted teaching practices, as teachers are often found to teach-to-the test (Barksdale-Ladd & Thomas, 2000). In more recent times the marking of examinations has also become highly controversial. No more so than in the United Kingdom where the delayed marking of Key Stage 3 (KS3) examinations led to a Parliamentary inquiry and ultimately the closure of the National Assessment Agency (NAA). Nevertheless, key government agencies, such as the Qualifications and Curriculum Authority (QCA) have recognised the importance of examination marking and commissioned a number of studies to investigate marking accuracy across a number of national curriculum tests (see QCA, 2009). This paper reports on one such study on the marking of KS3 mathematics (Baker et al., 2006). The main purpose of this paper is to report on cultural differences in mathematics test marking. It also reports on marking accuracy from a cross-cultural perspective (Australian v English markers) and the use of different marking locations (Home v Centre marking).

There has been a multitude of international studies investigating cross-cultural differences in mathematics performance and the underlying factors. Of great significance has been the ongoing international comparative research conducted by the various TIMSS and the PISA studies, which have investigated a wide range of issues related to the teaching and learning of mathematics (see Clarke, 2003). In these studies marking accuracy has not been a focus of investigation but a major technical concern. In the original TIMMS study, the designers recognised the need for marking (scoring) validity on free response items (Beaton et al., 1996). Hence scoring rubrics were developed and training sessions conducted to achieve a high rate of reliability. As part of this process studies were completed that compared markers on common scripts both within and between countries. For later TIMSS studies as well as PISA 2000 and 2003 similar strategies were used (see OECD, 2005). It was found that high rates of consistency could be achieved within and across the different countries that participated provided extensive training was provided. However, there were some variations found particularly on specific items. Although these studies have reported these findings, the underlying reasons behind these variations have not been investigated because the researchers' main aim was to achieve marking consistency.

This present study aimed to explore potential cultural reasons why there might differences in marking accuracy other than human error. To achieve this, Australian mathematics markers, chosen because of similarities between the two countries, were required to mark a sample of the 2006 KS3 mathematics exam (already marked and graded) under the same training conditions as their English counterparts. Their marking accuracy was then compared with the English markers. Furthermore as a secondary comparison, the markers were divided into two groups where one group marked at home and a second group marked at a common centre. To collect information on why marking differences might occur, the Australian markers kept diaries that documented their experiences and perceptions about the marking process.

Method

Participants

The study included 38 participants (F=21 and M=17) with an average age of 51.3 years from Sydney (Australia) or neighbouring towns. Thirty-two were currently teaching mathematics in a high school, three were recently retired from school teaching, and three taught mathematics at the tertiary level. Overall teaching experience in secondary schools was 26.5 years, with a mean marking experience of state examinations of 10.3 years. The participants were recruited because they were highly experienced and were paid market rates for state examination marking. Markers were randomly assigned to either the Home or Centre group.

Instruments and study procedures

The examination papers marked were copies of the 2006 KS3 mathematics tests, which had already been marked in England. All examination answer booklets were "cleaned" to ensure that no English markers' comments or marks were visible to Australian markers. A survey instrument collecting marker background information, such as demographic data and prior related marking experience, was developed.

Pre-training day materials

Prior to the training day, participants were mailed the Teacher Pack (QCA, 2006) that was identical to the pre-training package sent to the mathematics markers in England for the 2006 KS3 mathematics exam. The Australian markers were required to mark the five training scripts using the marking schemes prior to the training day.

The training day

Both groups undertook a 4-hour training session. During this time, an English Marking Trainer worked through the marking schedule using the 5 training scripts. On specific answers the trainer would emphasise certain aspects of the marking scheme, and explain why the correct mark should be 1 rather than 0, and vice versa. Such points often led to a discussion between the markers and the trainer. The meetings concluded by the markers filling in the survey and being given the following handouts, with instructions: a) a sheet to record the times spent marking and the number of scripts marked during these times; b) a marking journal, consisting of a number of sheets to record their thoughts as they marked; c) The *Edexcel* 2006 Key Stage 3 Mathematics External Marking: Commentary for Training Scripts

The following set of instructions for marking the box of 90 scripts was also provided: a) the scripts will be marked in Batches of 10; b) for the first batch of ten

you are required to sort the 10 scripts into separate bundles of the four tiers with the lowest tier first. For each tier you mark each question separately (E.g., all questions 1 in the tier, all questions 2 in the tier, and so on); c) For each remaining batch of ten, mark all papers 1's in the batch of ten in the exact order of the given sequence, mark all papers 2's in the batch of ten in the exact order of the given sequence, mark all mental maths papers in the batch of ten in the exact order of the given sequence; d) bundle each batch of ten with the rubber bands keeping the same order within the bundles and keep the bundles in the correct overall order. These marking instructions were provided to ensure a common approach to marking and that the correct order of script marking was maintained. In order to follow the English marking procedures as closely as possible and also to control some of these variables, it was decided to insist that the first 10 scripts were to be marked by sorting into tiers and then marking question by question. In this way, the first 10 scripts would act as further training and consolidate marker familiarity with the marking scheme.

Home and centre marking

At the conclusion of the training day, home markers took away their scripts. They were instructed to return them within 11 days, after having completed all 90 scripts or marked for a maximum of 20 hours. They were also instructed to contact the English Marking Trainer if they needed any advice in interpreting the marking scheme. Centre markers did not take their scripts home. For each day they attended, the relevant boxes were given out to the markers who took them to the designated marking room. At the end of each session, each completed bundle of ten scripts was collected and stored in a separate area. For each of the scheduled sessions, all the markers marked in one room, where the English Marking Trainer was present to answer any enquiries.

Results

Marking accuracy

In terms of marking accuracy there was no significant difference between the Central and Home markers. Furthermore, no differences were found between the Australian and English markers. However, marking accuracy was not the main focus of this paper and is reported elsewhere in more detail (see Baker et al., 2006).

Marking rates

The mean number of student scripts marked was 85.3 for the Home markers and 79.9 for the Centre markers. Mean times spent marking were 22.3 hours for the Home group and 19.0 hours for Centre markers giving mean marking rates of 3.9 and 4.2 respectively. Under a t-test there was no significant difference in marking speeds between the two groups, $t(35) = 1.27$, $p = 0.21$.

Patterns of marking in the Home group

Whereas the Centre markers were highly regulated in their marking patterns (Five 4-hour sessions), the Home markers were very much left to devise their own routines. Analysis of their Marking records revealed the following statistics. The group completed an average of 12.6 separate marking sessions: the highest being 24 and the least 7. On average a session lasted 1.9 hours. Whereas some markers were highly systematic in their approach (e.g. 11 sessions of 2 hours), others were far more irregular (e.g. 0.25, 4.25, 2, 1, 0.5, 0.75, 0.5, 1, 0.75, 1, 1.75, 2.75, 5.25 and 1 hour).

Generally Centre markers preferred to mark over small periods of time with substantial breaks.

Marking Journals Themes

Thirteen centre markers and 17 home markers returned their completed journals. These data were analysed using a simple content analysis technique. From each marker's report, a list of raw concepts (issues) was identified. If more than one marker documented the same issue then a frequency count was kept. Concepts with a common theme were then grouped together to form a category. For example, markers identified the 'tomatoes' and 'red kites' questions as difficult to mark. As a result both these questions were classified as members of the category named—Factors influencing speed and accuracy. Altogether, seven categories were identified: Training, Difficulties Interpreting the marking scheme, Positive aspects of the marking scheme, Factors influencing speed and accuracy, Comments on student performance, Increases in marking speed and Recommendations.

Training

A number of markers (n = 10) were concerned that they did not receive enough training, particularly home markers (n = 8), often referred to as 'Domestic' markers. One marker reported that too much time was spent arguing the pros and cons of the marking scheme rather than accepting it and moving forward. Four markers also expressed the view that having no quality control was disconcerting, affecting confidence and marking accuracy. Three observed that there was value in markers getting together to discuss the marking scheme, and one suggested that as a consequence the Centre markers would be better off. Several markers commented that reporting to a Senior Marker would have been beneficial, rather than working un-supervised. The following are some of the comments made:

> In Australia, 'domestic' markers still have the opportunity for clarification by discussing with their senior marker or with other team member via phone. Also, (there is) opportunity for this as you return boxes to the SM. In this study we were instructed not to communicate so some inconsistencies in applying the marking scheme are likely. (H11)
>
> There doesn't appear to be any quality control of what is happening with Home markers. Consequently I could mark very quickly but not accurately. (H13)
>
> Explanations that are "grey" – would be easy to get a "ruling" if senior markers there – can seem trivial to ring about one little thing. E.g. Is diagonal of enlarged square root (100) or root (200) student said " because" root (100) would show that the diagonal line and the side are the same length" seems reasonable but doesn't fit the manual. Will ring when I have another query as well. (H14)
>
> A longer briefing session or 2nd sample marking with a follow up meeting to discuss differences would have been beneficial. (H16)
>
> The process is very different as we are not being check-marked along the way – feedback as to how consistent I am marking would be helpful, although not really necessary as this marking does not effect the student's mark on the paper. (C8)

Difficulties Interpreting the Marking Scheme

A number of issues were reported in this category. Five markers reported that the marking scheme was far too complex and could have been simplified, with four reporting that summary information would have been useful. H16 commented in detail on this:

> Sometimes there is a lot of "jargon" to sift through in the "correct response" / additional guidance within. For example, Paper 1 p24 / 25, there is a lot in their 2 pages and if you want markers to keep speed up, maybe these guidelines need to be more concise / with a "tone" explained by what you are looking for. Checking up on some of the detail slows you down. We usually mark via 1) the solution and 2) rubric of intent from the people who set the exam not essays to be read through.

Eight markers believed that the scheme often penalised students who did not fit the expected answers and this made such scripts difficult to mark. Markers often reported that they felt students should have got a mark for correct working, but were unable to give it. For example H10 reports, "All the working is shown for Paper 1, Q1, but because no addition sign, therefore zero". Other similar comments on this theme included over-pedantic penalisation of marks for currency (£7.1, not given a mark), incomplete processing (7/2, no marks), substituting the wrong decimal symbol and probability (7 out of 10, no marks). Three markers thought that the scheme was inconsistent overall, sometimes ignoring an error that they had previously penalised. A possible explanation for this perception was the different emphases placed on conceptual and processing errors on specific questions. Three markers reported finding the idea that correct processing can be worth no marks given certain conceptual errors disturbing, as the following comments illustrate:

> I find the notion of "conceptual" errors a new one - one which was hard to stomach in some questions where a student lost 2 marks even though they knew the process. E.g. Q11 Paper 1 tier 3 – 5 (H11)

> Students drilled will more than likely leave out the "carry 7" How often do you write down your "carries" I never do. We penalise the better students here in the "design" of the question (i.e. the working is in the question). Do they know that they lose marks if they don't write the "7"? Do they know to show all the multiples etc? (H16)

There were other clashes in 'culture' reported. For example, three reported some difficulty with penalising a subsequent error. For example, if a student gave an answer as 21/4 and then incorrectly simplified to 5.1 this would be marked correct at the Higher School Certificate (HSC). It is worth noting that for HSC marking, there are acronyms that are widely used to express this situation, either CNE (correct numerical expression) and/or ISE (ignore subsequent error), although it should be pointed out that the HSC involves far more complex problems for older students. It was often voiced in personal anecdotes but not widely reported in the journals, that many believed that the English system looked to take away marks, rather than reward them. One marker reported that it was difficult following the marking scheme because of the underlying 'foreign' philosophy, another commented:

> The method of marking seems at odds with the way we mark in NSW e.g. marking complete paper as compared to marking a question, we look to give marks rather than deduct. (C10)

Three markers argued that the emphasis seemed to be on the final answer rather than the working, although a counter example reported by one marker was that the simultaneous equations question required working for a mark even though a correct answer might be given. Two markers reported that there were such subtle differences in some answers that they wanted to be free to make more judgements themselves, as the following quote illustrates:

> There was just some very subtle difference in some answers that were awarded 0 or 1 mark i.e. "multiples of 4 are even" and "4 is an even number" needed to use 'judgement flags' and would have liked to have the option where I felt the student

> was more correct than wrong – but their option was not on the answer sheet. (H17)

Finally, again illustrating the clash in cultures, two markers reported that a difficult aspect of the marking scheme was that it was predetermined and they had no part in its development and consequently no ownership.

Positive Aspects of the Marking Scheme

In contrast to the difficulties outlined above, there were some positive aspects of the marking scheme recorded, although not as many. However, six (5 home markers) reported that they found the comprehensive nature of the scheme very helpful. Two such comments were as follows:

> The mark scheme for all the papers was excellent. I liked the "minimally acceptable explanations" and the "incomplete or incorrect explanation." It certainly made the marking scheme easy to understand. (H6)

> Marked question by question (tiers 3 – 5) slow and laborious but starting to get a handle on the marking scheme (in particular the detailed guidelines given). I can see the need for the extra detail – particularly as a home marker. (H19)

Factors Influencing Speed and Accuracy

Responses on this theme were very similar. Twelve markers stated that there were far too many parts (questions per examination question, examination papers and tiers) to mark, interfering with their ability to remember the scheme and their marking confidence. Seven pointed out that the sheer number of pages to be turned was detrimental to speedy and accurate marking. Although not reported in the journals, anecdotal evidence suggested that an additional factor was the stapling of the scripts together, which made the flipping from page-to-page cumbersome. In contrast, the original student booklets were more user-friendly. Furthermore, four reported that if they discovered one of their own errors, it was again highly time-consuming to look back over previously marked scripts. Typical comments were as follows:

> Very slow having to turn pages so often – impossible to keep all the marking schemes in my head as we do in HSC marking. (H3)

> The marking of different papers and different tiers meant that I couldn't get into a flow of marking. Consequently I feel that I've probably made many more mistakes than I am used to. (H13)

> Remembering such a huge number of options makes marking incredibly slow. I also found it difficult to remember the tables and was slowed up by having to check continually. (C1)

Another major factor to influence speed and accuracy was the number of questions that required written explanations. Fifteen markers made this point directly or nominated particular questions that were most problematical. Chief amongst these were the questions labelled 'red kites' (n = 9), 'tomatoes' (n = 8) and 'odd or even?' (n = 4). Three markers added that they were very different types of questions to what they were used to. In addition, some non-explanation problems were also nominated as difficult to mark, such as 'five cubes' (n = 5), which required a number of 3-D perspectives.

Further, some markers (n = 4) reported that it was difficult to mark Tier 6-8 because they appeared so infrequently, particularly later questions, which many students did not try, and therefore they did not have sufficient exposure to get used to them. Finally, a major detriment to marking was the readability of the student scripts.

Eleven markers reported that there were many cases of poor photocopying/scanning which made it difficult or impossible to read. Two such comments were as follows:

> The indication that the question is worth 2 or 3 marks is good. Unfortunately, the extra line for marks was often missing due to the scanning process. (C1)
>
> MA803 Paper 1 Q8 Matching – too many crossed out – may have been different colours on the original, impossible to mark. (H10)

It was also reported that sometimes the markers guessed the student response, or simply left that question unmarked.

Comments on Student Performance

A number of observations were made about student performance on the examination. Generally, comments were not flattering. For example, individual markers reported that students were not good at the number plane, indices, simplifying fractions, Pythagorean theorem, graphs and circle geometry. More consensus was reached on algebra (n = 3), although one marker qualified this assessment by stating except Tier 6-8 students. Four markers highlighted English communication skills as a problem. Two made the assessment that some students were entered in too higher tiers, and two thought the standard overall was poor. In contrast, individual markers expressed the view that students were used to explanation questions, they were better at some content than others, they were very good at Q20, Paper 1 (Tier 3-5), and one reported enjoying students answers. Other observations included identifying common student errors on questions involving 'powers of 5' and 'three dice'.

Marker recommendations

A few markers made some recommendations mostly connected to splitting the marking into parts (n = 3) or marking only one tier (n = 3). One marker reported that it would be better to mark all of one tier first, before starting the next one. One marker also observed that the present format would be best marked corporately (at a centre), while one believed that a multiple-choice section would effectively cater for a number of the questions asked.

Conclusions

The information recorded in the journals indicated that the markers had experienced some difficulties on this task. Many felt, particularly home markers, that they did not initially receive enough training. Furthermore, a lack of further training or quality control had impacted on their confidence and ability to become adjusted to the marking scheme. Clearly, the task of marking whole exams, across 4 tiers, was very alien to their own marking experiences. The sheer number of pages to be turned in both student scripts and the marking schemes impacted on their speed and accuracy, as well as the sheer number of parts to become familiar with. Furthermore the Australian teachers reported that they would have liked more involvement in the development of the actual marking schemes, as 'ownership' was considered important. These findings indicate that the marking methods used by the KS3 markers may not be best practice.

Aspects of the marking scheme also took some adjustment. Different emphases on calculation accuracy, mathematical processes and conceptual understanding often created the perception that the marking scheme was inconsistent overall. Furthermore a high number of explain-type questions was also culturally different to their own experiences and led to difficulties in awarding marks. Other

positive aspects included the recognition by some that the marking scheme was very thorough and that some of the questions were creative and useful to the markers' own teaching practices. However, despite many reservations it was found that experienced Australian mathematics markers could adapt and successfully mark KS3 exams.

Finally this study may have some implications for the broader theme of assessment. Much of the research into assessment has focused on the importance of formative assessment and task selection, and not how such tasks are marked. It is clear that some items in the KS3 test were designed to test more than just facts. Yet the Australians had some difficulty marking particularly questions that emphasised conceptual processes despite being familiar with standards-based marking and formative assessment. Their methods of marking such questions were different, and they often felt the English method was more punitive. Here was a clear cultural difference where both countries could possibly learn from each other. It also suggests that providing good assessment tasks need to be backed up by equally good marking schemes in order to provide appropriate feedback. It is therefore suggested that further cross-cultural research into marking may be useful in supporting the bigger picture of assessment reform. As Clarke (2003) observed, assessment itself warrants more international comparative research.

References

Baker E. L., Ayres P, O'Neil H, Choi K, Tettey M and Sylvester R 2006. *KS3 Mathematics Marker Study in Australia: Report to the National Assessment Agency of England*. University of Southern California: Sherman Oaks, CA.

Beaton, A., Mullis, I., Martin, M., Gonzalez, E., Kelly, D., & Smith, T. 1996. *Mathematics Achievement in the Middle School Years: IEA's TIMSS*. : Chestnut Hill, MA, USA: Center for the Testing, Evaluation, and Educational Policy, Boston College.

Barksdale-Ladd, M.A. & Thomas, K.F. 2000. What's at stake in high-stakes testing. *Journal of Teacher Education,* 51: 384-397.

Clarke, D. 2003. International comparative research in mathematics education. In Second International Handbook of Mathematics Education, ed A. J. Bishop, M.A. Clements, C. Keitel, J. Kilpatrick and F. K.S. Leung, 143-184. Dordrecht: Kluwer Academic Publishers.

Organization for Economic Co-operation and Development 2005. *PISA 2003 Technical Report*. Paris: OECD

Qualifications and Curriculum Authority (2006). *The KS3 (Levels 3-8) Mathematics Test: Teacher Pack*. London: QCA.

Qualifications and Curriculum Authority 2009. *Research into marking quality: Studies to inform future work on nation curriculum assessments*. London: QCA.

Acknowledgements

The author wishes to acknowledge the contribution made by Eva Baker, Harry O'Neil, KC Choi, Margaret Tetty, Roxanne Sylvester and Colin Watson to this article. The author also wishes to thank Advance Design Information (Sherman Oaks, California) and the National Assessment Agency for funding the study.