CAMBRIDGE ASSESSMENT

The 'Marking Expertise' projects:

Empirical investigations of some popular assumptions

Irenka Suto and Rita Nádas
Research Division
Cambridge Assessment

Paper to be presented at the annual conference of the
International Association for Educational Assessment,
Baku, Azerbaijan, 16th - 21st September, 2007.

Contact details
Dr Irenka Suto
Research Division
Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU

+44 1223 553855
suto.i@cambridgeassessment.org.uk

This paper is also presented in *Research Matters: A Cambridge Assessment Publication*, 4 2-5.

Abstract

Cambridge Assessment continually evaluates the quality of its assessments, and the Research Division is conducting a series of inter-related studies investigating factors that could affect the accuracy of examination marking. These factors may include markers' teaching and marking experience and the depth of their subject knowledge. Certain personality traits may also be important. Furthermore, different types of question have been found to place different demands on markers. Our overall aim is to establish which skills and experiences are necessary for marking which question types.

The research explores diverse questions from UK and International GCSEs, taken from past examination papers. For each question, responses from a mixture of candidates were selected and all marks and annotations were removed. Markers with different backgrounds undertook some training and re-marked the responses. Additionally, Kelly's Repertory Grid technique was used with the most senior examiners to elucidate question features associated with accuracy among different marker groups.

Our initial studies focussed on maths and physics and a number of key findings from this research will be presented. More sophisticated follow-on studies are currently underway, investigating biology and business studies; these will also be outlined.

Introduction

Recent transformations in professional marking practice, including moves to mark some examination papers on screen, have raised important questions about the demands and expertise that the marking process entails. What makes some questions harder to mark accurately than others, and how much does marking accuracy vary among individuals with different backgrounds and experiences? It is becoming increasingly feasible for questions to be marked on a question-by-question basis by diverse groups of markers. While the differences between marking multiple-choice questions and long essays may seem axiomatic, an evidence-based rationale is needed for assigning questions with more subtle differences to different marker groups. We are therefore conducting a series of interrelated studies, exploring variations in accuracy and expertise in GCSE[1] examination marking.

In our first two linked studies, collectively known as *Marking Expertise Project 1*, we investigated marking on selected GCSE maths and physics questions from June 2005 examination papers administered by Oxford Cambridge and RSA Examinations (OCR) . Our next two linked studies, which comprise *Marking Expertise Project 2*, are currently underway and involve examinations from both Cambridge International Examinations (CIE) and OCR. This time we are focussing on International (I) GCSE biology questions from November 2005 and GCSE business studies questions from June 2006.

All four studies sit within a conceptual framework in which we have proposed a number of factors that might contribute to accurate marking. For any particular GCSE examination question, accuracy can be maximised through increasing the marker's personal expertise and/or through decreasing the demands of the marking task, and most relevant factors

[1] General Certificates in Secondary Education (GCSEs) are the first formal qualifications that many people in England and Wales obtain, usually at the end of compulsory education (aged sixteen). The responsibility for administering and awarding GCSEs is held by independent Awarding Bodies, who appoint professional examiners to mark candidates' examination scripts according to detailed and carefully structured mark schemes. International (I) GCSEs are taken by candidates in over a hundred countries worldwide.

can be grouped according to which of these two routes they contribute to. For example, personal expertise might be affected by an individual's subject knowledge, general knowledge, education, marker training (Shohamy *et al*., 1992; Powers and Kubota, 1998; Royal-Dawson, 2005), personality (Branthwaite *et al*., 1981; Greatorex and Bell, 2004; Meadows, 2006), teaching experience, and marking experience (Weigle, 1999), as well as knowledge of how best to utilise different marking strategies (for discussion of such strategies, see Suto and Greatorex, 2006, *in press*). Task demands, on the other hand, might be influenced by a question's length and features, the complexity and unusualness of a candidate's response, complexity of the cognitive strategies needed to mark it, and the detail and clarity of the accompanying mark scheme (Coffman and Kurfman, 1966; Raikes and Massey, 2007). A lot of these factors are popularly assumed to play a role in accuracy, yet research in the area is relatively sparse.

In this article, we present a summary of some key aspects and findings of the two studies comprising our first project. This research is described in depth elsewhere (Greatorex *et al*., 2007; Nadas and Suto, 2007; Suto and Nadas, 2007, *in press*). We end the article by looking ahead to our second project on marking expertise, which is currently in progress.

Marking Expertise Project 1: Study 1

Aim

The main aim of our first study was to explore potential differences in marking accuracy between two types of maths and physics markers: 'experts' and 'graduates'. Experts differed from graduates in that they had professional experience of both teaching and marking examinations, whereas graduates had neither teaching nor marking experience; however, all the markers had a relevant bachelor's degree. Further aims of the study were:

1. to explore the potential effects and interactions of two other key factors that may affect marking accuracy:

    a. intended question difficulty (for the candidate) within examination papers, as indicated by the tier(s) of the examination paper on which questions appeared
    b. the complexity of the marking strategies apparently needed to mark different questions within examination papers
2. to investigate individual differences in accuracy among markers
3. to explore the effects of a standardisation meeting (in which all markers reviewed and discussed their marking with their Principal Examiner) on accuracy
4. to explore potential relationships between marking accuracy and
    a. marking times
    b. self-confidence in marking
    c. perceived understanding of the mark scheme.

Design

For both subjects, groups of expert and graduate markers were led by a Principal Examiner in the marking of identical samples of candidates' responses on a question-by-question basis. Several brief questionnaires were also completed by all markers, which included questions about their self-confidence about their marking. A quantitative analysis of the data was then conducted, utilising three different measures of accuracy: P_0 (the overall proportion of raw agreement between two markers), Mean Actual Difference (an indication of whether the marker is on average more stringent or more lenient than his or her Principal Examiner), and Mean Absolute Difference (an indication of the average magnitude of mark differences between the marker and his or her Principal Examiner).

Key findings

All three measures of accuracy generated similar results, and the study yielded several interesting findings:

- There were very few significant differences in the accuracy levels of experts and graduates for either subject. For mathematics, the marker groups differed significantly (i.e. at the 5 % level) on just one question out of twenty. For physics, the marker groups differed significantly on two questions out of thirteen. In all cases, the differences in accuracy were small.

- For both subjects, accuracy in general (among all markers) was found to be related to intended question difficulty. Broadly speaking, questions that appeared on higher tiers (and were therefore intended to be harder for candidates) were harder to mark.

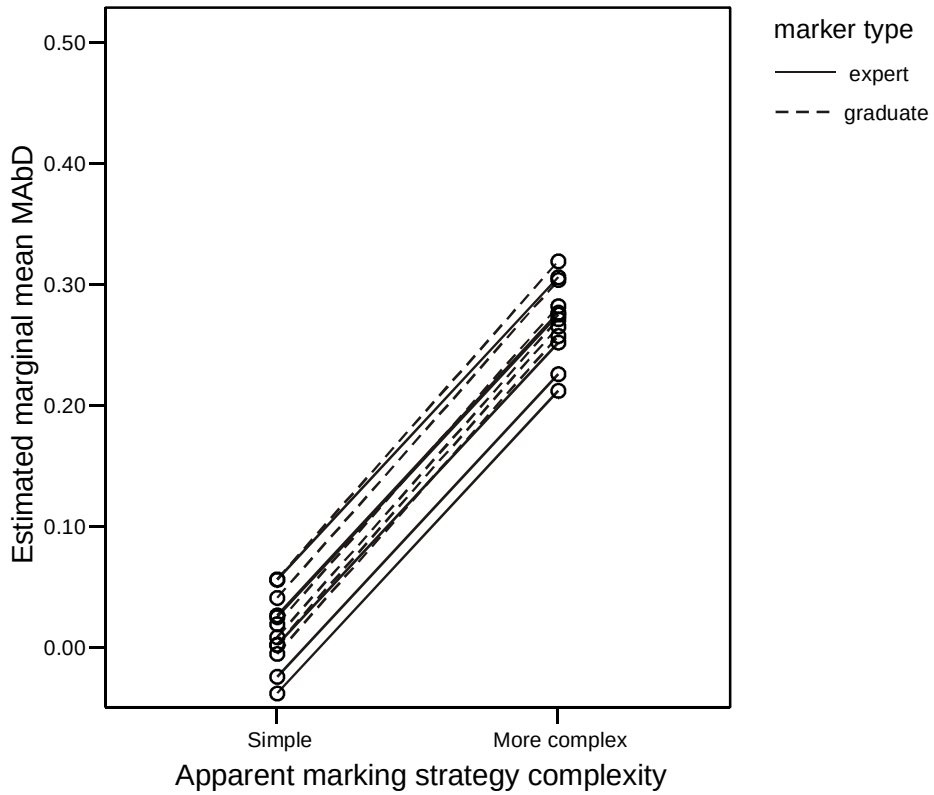
- For both subjects, accuracy in general (among all markers) was found to be related to apparent cognitive marking strategy usage. Broadly speaking, questions judged by the researchers to entail only simple strategies (matching, scanning for simple items) were marked more accurately than were those judged to entail more complex strategies (scanning for complex items, evaluating, and scrutinising) instead of or in addition to simple strategies.

- For both subjects, the factors of intended question difficulty and apparent marking strategy were found to interact. That is, the effect of apparent strategy usage on how accurately a question was marked depended in part upon that question's intended difficulty for candidates.

- For physics in particular, there were significant *individual* marker differences in accuracy. Moreover, in physics there was a strong relationship between

individuals' accuracies on questions requiring only apparently simple marking strategies and their accuracies on questions requiring apparently more complex marking strategies. Figure 1 illustrates this finding for the analysis of Mean Absolute Differences (MAbD): the lines representing individual markers are almost all parallel to one another and there is little overlap.

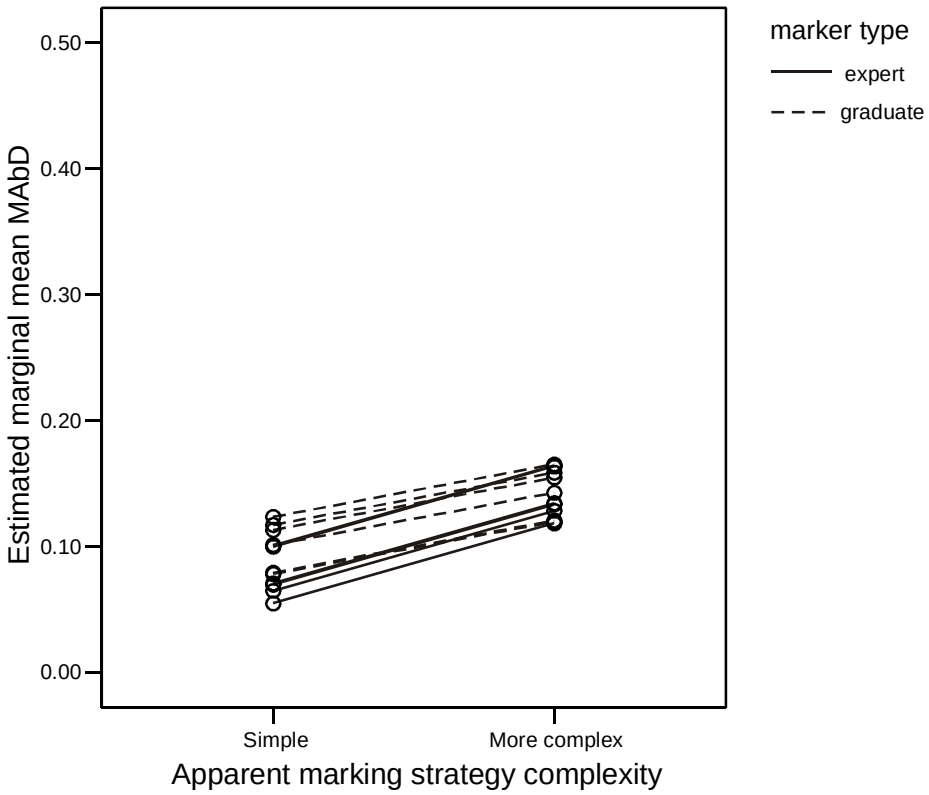
Figure 1: Graph showing estimated marginal mean MAbD values for individual physics markers (experts and graduates) for questions with different apparent marking strategy complexities



- In contrast, there was no distinctive *overall* relationship of this kind for maths. However, the within-group differences in the accuracies with which simple strategy and more complex strategy questions were marked were smaller than the between-group differences. This is shown in Figure 2: the lines representing individual experts are all of a similar gradient, and the lines representing

graduates are all of a different gradient. This suggests that the two marker groups may have had distinct marking behaviours, even though *overall*, they did not differ significantly in their marking accuracy. This issue may be worthy of investigation in a larger study.

Figure 2: Graph showing estimated marginal mean MAbD values for individual maths markers (experts and graduates) for questions with different apparent marking strategy complexities



- For both subjects, the standardisation meetings were effective in bringing the two marker groups closer together in their marking. When the meetings' effects were considered for each marker type separately, they were found to have been much greater on graduates than on experts. Overall the meetings had positive effects on accuracy for experts in physics, and for graduates in both subjects.

- For both subjects, the largest post-standardisation meeting improvement in accuracy arose when graduates marked questions requiring apparently more complex marking strategies. However, this is also where there was the most potential for improvement.

- For both subjects, there were no striking relationships between self-reported marking times and accuracy.

- For maths, experts were more self-confident in their marking than graduates were. However self-confidence ratings were not related to actual marking accuracies for either group.

- Conversely, for physics, there were no differences in the self-confidence (in marking) of experts and graduates. Experts' self-confidence ratings after marking the main sample of candidate responses correlated positively with their actual marking accuracies, whereas for graduates there was a negative correlation.

- For both subjects, there were no striking relationships between perceived understanding of the mark scheme and marking accuracy.

Conclusions

We drew a number of conclusions and implications from our first study:

- When led by a Principal Examiner and having attended a standardisation meeting, some graduate maths and physics markers mark almost all questions as accurately as their expert counterparts can. It appears that the awarding bodies could potentially look towards relaxing some of their guidelines for recruiting maths and physics examiners to mark at least *some* types of questions. However, further research is clearly needed. In other subjects, differences in the accuracies of experts and graduates may exist.

- There are grounds for allocating higher tier questions (that are intended to be hardest for candidates) and the questions that entail apparently more complex marking strategies to whichever examiners are ultimately considered to be the 'best' markers. Although there may be no real distinction between the accuracy of graduates and experts for GCSE maths and physics marking, further research could reveal differences in accuracy among other marker types, for example those with only A-level or GCSE subject knowledge.

- The striking relationship between apparent marking strategy complexity and marking accuracy provides a further validation of Cambridge Assessment's earlier research on cognitive marking strategies (Suto and Greatorex, 2006, *in press*); the distinction between apparently simple and apparently more complex marking strategies is clearly meaningful, as it can contribute usefully to predictions of how accurately particular questions will be marked.

- As Figure 1 indicates, if a physics marker's accuracy (either expert or graduate) on apparently simple physics questions were known prior to the 'live' marking of apparently more complex questions, then this could be used (for example, in a

screening procedure) to predict the likelihood of whether s/he would meet a pre-determined accuracy threshold for marking apparently more complex questions.

- The significant individual differences identified among physics markers could be due to personality characteristics; however, research in this area is needed.

- Future examination questions could be designed to avoid marking strategies and question features associated with lower accuracy. However, this would need to be handled very cautiously: effects on validity would need to be considered.

- The findings add weight to research literature extolling the importance of procedural training for inter-marker agreement. This is particularly important for graduates. It could be argued that standardisation meetings should focus almost exclusively on the questions entailing apparently more complex marking strategies.

- Broadly speaking, it appears likely that a marker's self-confidence in his or her marking is generally a poor predictor of accuracy, and markers have very limited understanding of their own marking expertise.

Marking Expertise Project 1: Study 2

Aim

The aim of our second study, which followed on directly from the first, was to identify question features that distinguish questions that are marked highly accurately from those marked less accurately. Having focussed on personal marking expertise in our first study, we were keen also to address the other half of the accuracy equation: the demands of the marking task.

Design

Differences among GCSE maths and physics questions with differing marking accuracies were explored qualitatively. To do this, we used Kelly's Repertory Grid (KRG) technique (for a full discussion of KRG, see Jankowicz, 2004) and semi-structured interview schedules in meetings with each of the two Principal Examiners who participated in Study 1. These methods enabled the PEs to identify ways in which questions differed from one another, and thereby formulate distinct question features or 'constructs'. The PEs then rated all questions on each construct using a scale of 1-5. (For dichotomous constructs, a yes/no rating was given instead). In an analysis of the construct data, possible relationships between each question feature and (i) marking accuracy, (ii) apparent cognitive marking strategy usage, and (iii) question difficulty (for the candidate) were then investigated.

Key findings

- For each subject, accuracy in general (among all markers) was indeed found to be related to various subject-specific question features (constructs). Some of these features were related to question difficulty and/or apparent marking strategy complexity. Others appeared to be related to accuracy only.

- For maths, it was concluded that four question features combine with question difficulty and apparent marking strategy complexity to influence marking accuracy. They are:
    - *Alternative answers*: the extent to which alternative answers are possible.
    - *Context*: the extent to which the question was contextualised.
    - *Follow-through*: whether follow-through marks are involved (i.e. marks that are contingent on the award of other marks within a question).
    - *Marking difficulty (PE's perception)*: the PE's personal estimation of how difficult the question is to mark.

However, the questions of if, and the extent to which, any of these factors interact with one another to affect marking accuracy, could not be answered definitively.

- For physics it was concluded that seven features may be useful in predicting marking accuracy together with question difficulty and apparent marking strategy complexity:
    - *Reading*: how much the candidate is required to read.
    - *Diagram*: the presence and importance of a diagram.
    - *Single letter*: whether single letter answers are required.
    - *Writing*: how much the candidate is required to write.
    - *MS flexibility*: whether the mark scheme offers a choice of responses or is absolutely inflexible
    - *Marking time:* how long the question takes to mark.
    - *Marking difficulty (PE's perception)*: the PE's personal estimation of how difficult the question is to mark.

  As with maths, however, the questions of if, and the extent to which, any of these factors interact with one another to affect marking accuracy, could not be answered.

Conclusion

The key conclusion from our second study was that the subject-specific question features (constructs) that are related to marking accuracy provide a rationale for allocating particular questions to different marker groups with different levels of expertise. However, there is a need for further research into the constructs' generalisability, involving other syllabuses and also other subjects.

Marking Expertise Project 2

At the start of the Marking Expertise Project 1, it was proposed that for a given GCSE examination question, accuracy can be improved either through increasing a marker's

expertise or through reducing the demands of the marking task, and that most other factors can be grouped according to which of these two routes they are most likely to contribute to. The project's findings (from both studies) fit comfortably within this framework. However, there were a number of limitations to the project. We explored only two examination subjects out of many, and for pragmatic reasons, we investigated only two types of marker: experts and graduates. Since experts had both teaching and marking experience and graduates had neither teaching nor marking experience, these two variables were not manipulated independently. Had there been any differences in accuracy between the two marker types, then the relative influences of marking experience and teaching experienced on accuracy could not have been ascertained.

We are seeking to address these issues in our second Marking Expertise project, which focuses on IGCSE biology and GCSE business studies marking. Again, we are exploring personal expertise and the demands of the marking task in two linked studies. However, in Study 1 of this second project, the participant group design is more sophisticated. For each subject, there are five participant groups, enabling us to investigate the relationships of four different variables with marking accuracy. The variables are:

- Relevant marking experience (i.e., experience of marking biology or business studies IGCSE or GCSE questions)
- Relevant teaching experience (i.e., experience of teaching GCSE biology or business studies)
- Subject knowledge (i.e. highest qualification in biology or business studies)
- General education (i.e. highest qualification in a subject other than biology or business studies).

The project will enable us to refine and develop our framework for understanding marking accuracy. We hope it will shed further light on the key question of how examination questions can best be assigned to markers with different sets of skills and experiences.

References

Branthwaite, A., Trueman, M. & Berrisford, T. (1981). Unreliability of marking: further evidence and a possible explanation. *Education Review* **33**, 1, 41-46.

Coffman, W.E. & Kurfman, D.G. (1966). *Single score versus multiple score reading of the American history Advanced Placement examination.* ETS Research Report no. RB-66-22.

Greatorex, J. & Bell, J.F. (2004). Does the gender of examiners influence their marking? *Research in Education* **71**, 25-36.

Greatorex, J., Nadas, R., Suto, W.M.I., & Bell, J.F. (2007). *Exploring how the cognitive strategies used to mark examination questions relate to the efficacy of examiner training.* Paper presented at the annual conference of the European Conference on Educational Research, Ghent, Belgium.

Jankowicz, D. (2004). *The easy guide to repertory grids*. Chichester: John Wiley & Sons.

Meadows, M. (2006). *Can we predict who will be a reliable marker?* Paper presented at the annual conference of the International Association for Educational Assessment, Singapore.

Nadas, R. & Suto, W.M.I. (2007) *An exploration of self-confidence and insight into marking accuracy among GCSE maths and physics markers.* Paper presented at the annual conference of the International Association for Educational Assessment, Baku, Azerbaijan.

Powers, D. & Kubota, M. (1998). *Qualifying essay readers for an Online Scoring Network (OSN).* ETS Research Report no. RR-98-20.

Raikes N. & Massey, A. (2007). Item-Level Examiner Agreement. *Research Matters: A Cambridge Assessment Publication* **4**, 34-37.

Royal-Dawson, L. (2005). *Is Teaching Experience a Necessary Condition for Markers of Key Stage 3 English?* Assessment and Qualifications Alliance report, commissioned by the Qualification and Curriculum Authority.

Shohamy, E., Gordon, C.M., & Kraemer, R. (1992). The Effects of Raters' Background and Training on the Reliability of Direct Writing Tests. *The Modern Language Journal* **76**, 27-33.

Suto, W.M.I. & Greatorex, J. (2006). A cognitive psychological exploration of the GCSE marking process. *Research Matters: A Cambridge Assessment Publication* **2**, 7-11.

Suto, W.M.I. & Greatorex, J. (*in press*). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process*. British Educational Research Journal.*

Suto, W.M.I. & Nadas, R. (2007)*. What makes some GCSE examination questions harder to mark accurately than others? An exploration of question features related to accuracy.* Paper presented at the annual conference of the British Educational Research Association, London, UK.

Suto, W.M.I. & Nadas, R. (*in press*). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education*.

Weigle, S.C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing* **6**, 2, 145-178.