CAMBRIDGE ASSESSMENT

Towards a new model of marking accuracy:
An investigation of IGCSE biology

Irenka Suto and Rita Nádas

Research Division
Cambridge Assessment

Paper to be presented at IAEA
September, 2008.

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

01223 553847
suto.i@cambridgeassessment.org.uk

Abstract

It is important that all public examinations should be marked highly accurately. In the UK and internationally, examinations for General Certificates in Secondary Education (GCSEs) influence the futures of thousands of candidates. However, what affects marking accuracy in GCSEs? Previously, we have conceptualised factors as contributing either to marking task demands, or to markers' personal expertise. An empirical study was conducted as a further elaboration of this work, investigating the relative roles of some key factors for a past International GCSE biology examination.

42 markers participated, comprising five groups: (i) experienced examiners, (ii) biology teachers, (iii) graduates in biology, (iv) graduates in other subjects, and (v) non-graduates. This design enabled the relative effects on accuracy of the following factors to be elicited: marking experience, teaching experience, highest education in a relevant subject, highest education in any subject, and gender. 23 examination questions were explored, varying in: format, number of marks, difficulty for candidates, and cognitive marking strategy complexity. All markers marked identical response samples for each question.

Logistic regression and ANOVA were used to model the accuracy data yielded, revealing education to be more important than experience. Our refined model may prove useful where evidence-based decisions surrounding marker recruitment and training are needed.

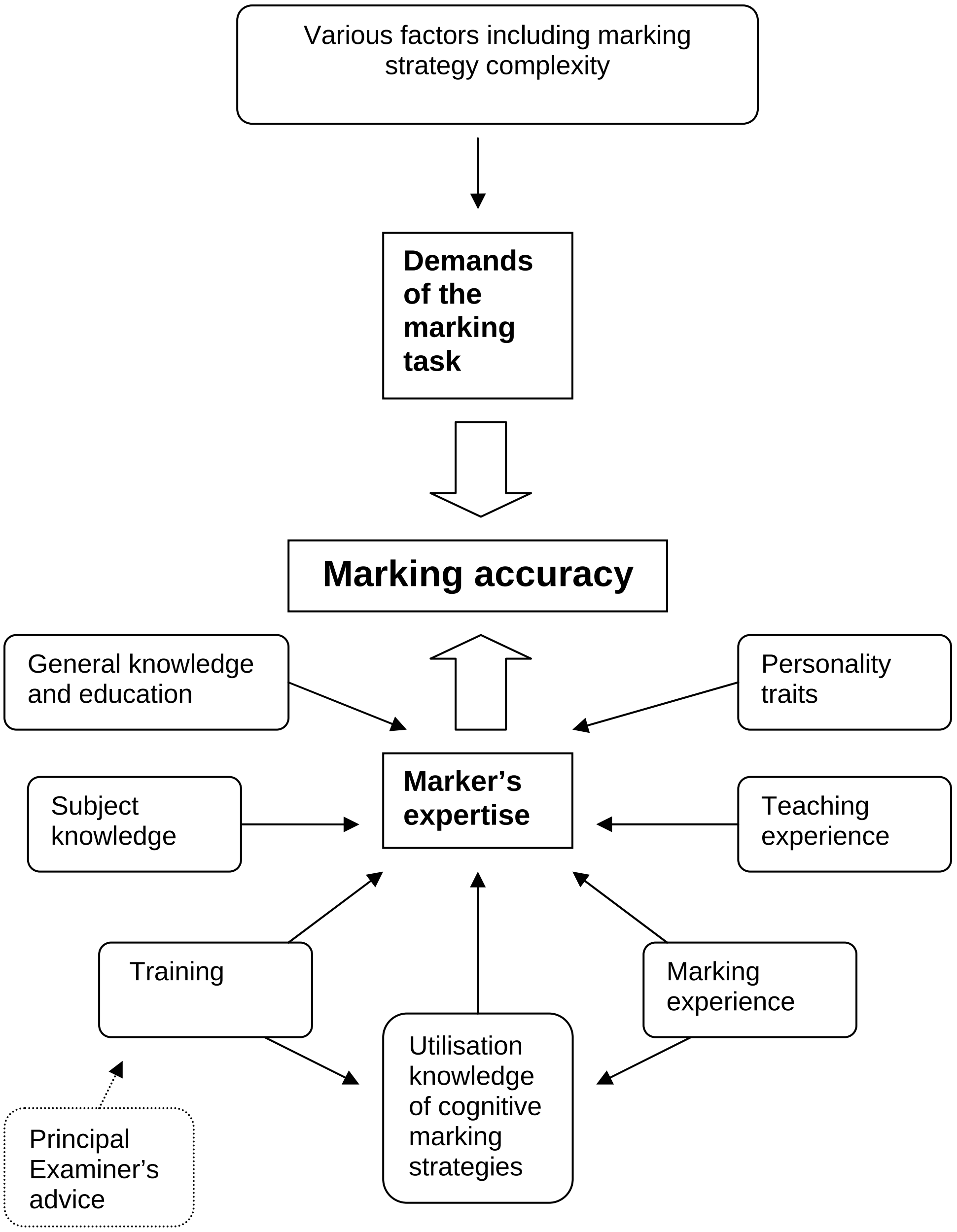
Background

The outcomes of public examinations often play pivotal roles in determining the directions that young people take at the end of compulsory schooling. For example, in the UK and internationally, examinations for General Certificates in Secondary Education (GCSEs) influence whether many thousands of candidates can proceed to further education or enter into employment. It is crucial, therefore, that public examinations are marked as accurately as possible, ensuring fair results for all. This necessity engenders a substantial and wide-ranging research question: what determines marking accuracy?

Within our research group, we have conducted a series of studies addressing this question and some of the issues that underpin it. Initial research focussed on the cognitive marking strategies used to mark GCSE and A-level examinations (Greatorex and Suto 2005, Suto and Greatorex, 2006, 2008, a, b). 'Think aloud' data was collected through empirical studies of experienced examiners, and was interpreted within a well-established dual processing theory of human judgement (Kahneman and Frederick, 2002). It was argued that seemingly basic marking strategies such as matching or scanning for small items of information (for example, individual letters or numbers in candidates' responses to examination questions) utilise simple *System 1* or *intuitive* judgmental processes. In contrast, more sophisticated marking strategies, such as scanning for larger items of information, and evaluating or scrutinising what candidates have written, utilise more complex *System 2* or *reflective* judgmental processes.

In subsequent research, we explored the relationship between the complexity of the cognitive marking strategies that examination questions entail and the accuracy of marking (Suto and Nádas 2007, 2008a, b). An empirical study was conducted in which past examination questions for GCSE mathematics and physics were explored, and the relationship was found to be strong: questions entailing only simple marking strategies were marked more accurately than those entailing more complex strategies. Through this research, a new theoretical framework for understanding marking accuracy was constructed. We conceptualised marking accuracy for a particular question as being affected by both (i) the demands of the marking task, including marking strategy complexity, and (ii) a marker's personal expertise. Arguably, accuracy can be improved both by reducing the demands of the marking task and by increasing a marker's personal expertise (see Figure 1).

Figure 1: Diagram summarising some key factors identified as likely to contribute to marking accuracy, indicating the main relationships hypothesised among them (adapted from Suto and Nádas, 2008 a)


Various factors including marking strategy complexity
Demands of the marking task
Marking accuracy
General knowledge and education
Personality traits
Subject knowledge
Marker's expertise
Teaching experience
Training
Marking experience
Utilisation knowledge of cognitive marking strategies
Principal Examiner's advice

Within the broader educational assessment community, it has long been established that in public examination marking in the UK, inter-marker agreement is imperfect, varying significantly among examination subjects as well as among teams of markers (Valentine, 1932; Murphy, 1978, 1982; Newton, 1996; Pinot de Moira, Massey, Baird and Morrissey, 2002; Laming, 2004). Unsurprisingly, therefore, the question of what contributes to a marker's personal expertise has been the focus of several recent studies. Various factors that potentially affect expertise in examination-marking have been investigated, including: marker training, (Baird, Greatorex & Bell, 2004; Royal-Dawson, 2005); marking and teaching experience (Suto and Nádas, 2008a); knowledge of the subject being examined (Meadows, 2006; Meadows and Billington, 2007; Meadows and Wheadon, 2007); gender (Greatorex and Bell, 2004); and personality traits (Branthwaite, Trueman and Berrisford, 1981; Meadows and Billington, 2007). This body of research draws from a variety of educational assessment contexts, and while it is far from definitive, it indicates that, with the exception of gender, all of the above factors do contribute in some way to marking expertise. However, while many factors appear to be important *per se*, their relative influences are far from clear.

Aims

The research summarised in this paper entailed a systematic investigation of five key factors identified in previous research as affecting, or likely to affect, personal expertise: marking experience, teaching experience, highest level of education in a relevant subject, highest level of education in any subject, and gender. An empirical study entailing experimental marking of past questions from an International GCSE (IGCSE) biology qualification was conducted, the aims of which were not only to confirm the importance of these factors, but moreover, to ascertain their relative roles. Knowing the importance of each factor relative to the others could contribute to awarding bodies' decisions about who is best placed to mark examinations of this kind.

Design and methods

An IGCSE qualification in biology from November 2005 was used in the research. It comprised many diverse question types, varying along several dimensions, including: format, number of marks, difficulty for candidates, and cognitive marking strategy complexity. A selection of 23 past examination questions was explored in the study.

Five groups of markers were asked to mark identical samples of candidates' responses on a question-by-question basis, using the relevant sections of the mark schemes designed for the original 'live' examination. The markers were led by an experienced Principal Examiner (PE), who was supported by a Team Leader (TL). The five marker groups were:

(i) '*Experts*' - experienced IGCSE examiners (N = 8)

(ii) '*Teachers*' - GCSE biology teachers with no marking experience (N = 9)

(iii) '*Relevant graduates*' - graduates in biology with no marking or teaching experience (N = 8)

(iv) '*Other graduates'* - graduates in other subjects with no marking or teaching experience (N = 9)

(v) '*Non-graduates'* - individuals with no university education and with no marking or teaching experience (N = 7).

The participant groups were designed so that the relative effects on accuracy of the following factors could subsequently be elicited: marking experience, teaching experience, highest level of education in a relevant subject, highest level of education in any subject, and gender.

Four samples of candidate responses were created:

- *Practice* sample: 5 different responses to each question
- *1st standardisation* sample: 10 different responses to each question
- *2nd standardisation* sample: 10 different responses to each question
- *Main* sample: 50 different responses to each question.

The *2nd standardisation* sample was marked only if a marker failed to reach the pre-agreed level of accuracy on the *1st standardisation* sample. Therefore, in total, each marker marked either 3 or 4 response samples for each question.

Analysis

The analysis summarised here focussed on the marking of the *main* response samples. To investigate marking accuracy, we chose to explore P_o, which is the proportion of raw agreement between the marks of each marker and the marks of the Principal Examiner. This is essentially a measure of how frequently two markers differ in their marking. P_o values were calculated for every marker for every question for the *main* response sample. ANOVA and logistic regression were used to model this accuracy data, thereby clarifying the relative roles of the factors under investigation.

Key findings

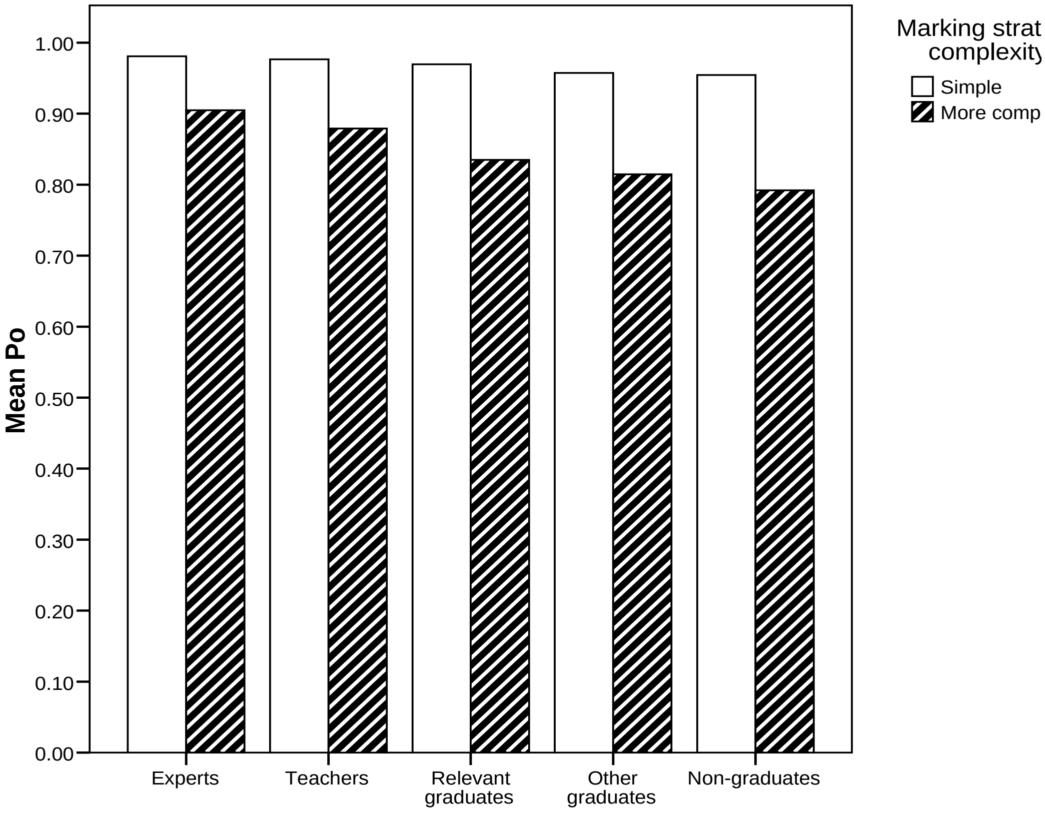
Marking accuracy was found to be generally very high, and differences among the five marking groups were all in the directions anticipated. The mean P_o values for the five marker groups are given in Table 1.

Table 1: Mean P_o values obtained across all questions on the main sample of candidate responses

Marker group	Mean P_o	s.d.
Experts	0.94	0.08
Teachers	0.92	0.09
Relevant graduates	0.89	0.12
Other graduates	0.88	0.13
Non-graduates	0.87	0.15

However, whilst the 13 questions entailing only *simple* cognitive marking strategies were marked with near perfect accuracy by all markers regardless of their backgrounds, the 10 questions entailing *more complex* marking strategies were marked *relatively* less accurately by all markers. These findings are illustrated in Figure 2. It can be seen that marking agreement with the PE was lowest among non-graduate markers when marking questions entailing *more complex* marking strategies.

Figure 2: Graph to show P_o values obtained by the five marker groups when marking the main sample of candidate responses to 'simple' and 'more complex' strategy questions



Note: A mean P_o of 1 would indicate perfect agreement with the PE on every question.

Given these findings, it was decided that the main analysis of factors affecting personal expertise should focus only on questions entailing *more complex* marking strategies. It utilised logistic models with effects analogous to AN(C)OVA for a linear model, for example:

$$\ln\left(\frac{P_0}{(1-P_0)}\right) = \alpha + \beta_i + \gamma_j + \delta_{ij}$$

where P_0 was the probability (proportion) of exact agreement between a marker and the PE, α was the constant, and β_i and γ_j were effects for factors (such as marker group and tier in Model 1) and δ_{ij} is the interaction between the two factors. Endpoint parametisation was used to set the models; i.e. the final question was set to 0, and statistics for all other questions were compared with those for the final question. The dependent variable was the agreement between a binary variable taking the value 1 when the marks on an individual question agreed, and the value 0 otherwise.

Several series of models of this form were run, with each model investigating at least one of the factors (potentially) affecting markers' expertise. Models were compared to ascertain how well they fitted the data, and the relative predictive values of the factors were thereby found to be (in order of importance):

(i) highest level of education in any subject
(ii) highest level of education in a relevant subject
(iii) teaching experience
(iv) marking experience
(v) gender (women marked more accurately than men did).

Marking experience was found to be associated (at present) with highest general education; that is, the contributions of these factors to marker expertise are not independent of one another.

Discussion

This study of examination marking in IGCSE biology has generated interesting findings pertaining to factors that contribute to marking accuracy. While the study is obviously limited in its scope and scale, its outcomes may prove most informative when interpreted together with those from previous studies. The finding that questions entailing *more complex* marking strategies are marked less accurately than those entailing only *simple* strategies corroborates previous research on these marking strategies (Greatorex, Suto and Nádas, 2008; Suto , Crisp and Greatorex, 2008; Suto and Nádas, 2008a, b). This adds further weight to arguments for using the strategies in the process of classifying examination questions in order to assign them to markers of differing expertise.

Furthermore, the finding that the level of a marker's highest education (either in general or in a relevant subject) is essentially a better predictor of accuracy than either teaching or marking experience is entirely in line with the outcome of our earlier study of GCSE mathematics and physics marking (Suto and Nádas, 2008 a). In the earlier study, it was found that graduates in relevant subjects but with neither teaching nor marking experience were able to mark as accurately as individuals with both teaching and marking experience. When taken together, the findings broadly suggest that when it comes to marking GCSE examinations in maths and science, education is more important than experience. Although the graph in Figure 1 may appear to indicate that marking and teaching experience are highly beneficial, the (current) association between marking experience and highest general education would account for much of the apparent benefit.

The finding that a marker's highest level of education in *any* subject is a better predictor of accuracy than his or her highest level of education in a *relevant* subject is open to a number of interpretations. The most likely of these is arguably that the key to successful marking is being able to follow marking instructions and interpret the mark scheme in the way its author intended. Generic academic abilities may be both necessary and sufficient for this, which may be common in the graduate population at large. Gaining experience in teaching and marking may help to strengthen and contextualise these abilities, but may still be less important overall. It is worth noting that ultimately, the level of marking accuracy deemed satisfactory for questions entailing *more complex* marking strategies is a matter of judgment, given that such questions entail an unquantifiable but inherent degree of subjectivity.

The finding that women marked more accurately than men did is surprising, given that Greatorex and Bell (2004) found no relationships between gender and GCSE marking in an earlier study. The reasons for this effect are unclear and require further investigation. As the Principal Examiner was male, it seems unlikely that his 'correct' marks (against which all other markers' marks were compared) favoured female marking styles in some respect.

Conclusions

Through the present study we have been able to refine the model developed in our previous research, of marking accuracy being maximised either through reducing marking task demands or through increasing personal expertise. For IGCSE biology, five of the factors affecting personal expertise can now be ranked in order of influence, and for maths and science subjects in general, it seems very likely that markers' level of education is more important influence on accuracy than experience.

It is intended that this research will prove useful to examination boards when evidence-based decisions surrounding marker recruitment are needed. At a time when rapid technological advances are influencing the development of policies for employing new types of marker, it is essential to ensure that all questions continue to be marked by individuals with appropriate background experience. Given the diversity of examination questions that appear in GCSE examinations, this is no simple task.

References

Baird, J-A., Greatorex, J. & Bell, J.F. (2004) What makes marking reliable? Experiments with UK examinations, *Assessment in Education* 11 (3) 333-347.

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

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

Greatorex, J. & Suto, W.M.I. (2005) *What goes through a marker's mind? Gaining theoretical insights into the A-level and GCSE marking process*. Paper presented at the 6th Annual Conference of the Association for Educational Assessment – Europe, Dublin, Republic of Ireland, November 2005.

Kahneman, D. & Frederick, S. (2002) Representativeness revisited: attribute substitution in intuitive judgement. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: the psychology of intuitive judgement*. Cambridge: Cambridge University Press.

Laming, D. (2004) *Human judgement: the eye of the beholder.* London: Thompson.

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

Meadows, M. & Billington, L. (2007) *The right attitude for marking?* Paper presented at the annual conference of the British Educational Research Association, 5th-8th September, London, UK.

Meadows, M. & Wheadon, C. (2007) *Selecting the conscientious marker – a study of marking reliability in GCSE English.* Paper presented at the annual conference of the International Association for Educational Assessment, Baku, Azerbaijan, 16-21 September.

Murphy, R.J.L. (1978) Reliability of marking in eight GCE examinations, *British Journal of Educational Psychology* 48, 196-200.

Murphy, R.J.L. (1982) A further report of investigations into the reliability of marking GCE examinations, *British Journal of Educational Psychology* 52, 58-63.

Newton, P. (1996) The reliability of marking General Certificate of Secondary Education scripts: mathematics and English, *British Educational Research Journal* 22, 404-420.

Pinot de Moira, A., Massey, C., Baird, J. & Morrissy, M. (2002) Marking consistency over time, *Research in Education* 67, 79-87.

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.*

Suto, W.M.I., Crisp, V. & Greatorex, J. (2008) Investigating the judgemental marking process: an overview of our recent research, *Research Matters: A Cambridge Assessment Publication* 5, 6-8.

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. and Greatorex, J. (2008 a) What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process, *British Educational Research Journal*. 34 (2) 213-233.

Suto, W. M. I. and Greatorex, J. (2008 b) A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations, *Assessment in Education: Principles, Policy and Practice* 15 (1) 73-90.

Suto, W.M.I. & Nádas, R. (2007) The ‘Marking Expertise’ projects: empirical investigations of some popular assumptions, *Research Matters: A Cambridge Assessment Publication 4, 2-5.*

Suto, W.M.I. & Nádas, R. (2008 a) What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers, *Research Papers in Education*.

Suto, W.M.I. & Nádas, R. (2008 b) Why are some GCSE examination questions harder to mark accurately than others? Using Kelly’s Repertory Grid Technique to identify relevant question features, *Research Papers in Education*.

Valentine, C.M. (1932) *The reliability of examinations: an enquiry with special reference to the entrance examinations to secondary schools, the school certificate examinations, and the award of scholarships at universities*. London: University of London Press.