**Issue 4** June 2007

# Research Matters

CAMBRIDGE ASSESSMENT

UNIVERSITY *of* CAMBRIDGE
Local Examinations Syndicate

# Research Matters : 4

A CAMBRIDGE ASSESSMENT PUBLICATION




If you would like to comment on any of the articles in this issue, please contact Sylvia Green.
Email:
researchprogrammes@cambridgeassessment.org.uk

The full issue and previous issues are available on our website:
www.cambridgeassessment.org.uk/research

## Foreword

As new technologies begin to emerge in assessment, it sometimes feels as if progress is being made on delivery mechanisms without commensurate development in understanding of measurement and related aspects of assessment processes. This fourth edition of *Research Matters* should be a corrective to that. It drills down deeply into marking. It extends understanding, through empirical work on marking processes; but in addition validates the research methods which can be deployed to undertake such work. This work is not only vital for developing a better understanding of contemporary assessment processes and for developing more refined procedures regarding management and quality assurance of marking, it is vital also for benchmarking the changes which will inevitably come with the introduction of new technologies.

There is a misplaced debate on 'will we use technology to do what we do now, only more efficiently, or will it be used to assess in quite new ways?' All the evidence we have to date suggests that even if we set out to use new technology simply to make existing processes more efficient, the introduction of new technologies – such as on-screen marking – always has some impact on the technical characteristics of assessment. The work outlined in this issue challenges assumptions and generates new evidence on marking; in doing this it additionally provides us with an invaluable reference point as we monitor the impact of structural change.

**Tim Oates** *Group Director, Assessment Research and Development*

## Editorial

The main themes of this issue relate to the psychology of marking, cognitive processes affecting accuracy, and issues related to quality assurance in marking processes. The first three articles focus on marking practices and the factors that impact on them.

In their article Suto and Nadas report on their work in the context of marking expertise, considering the demands and expertise that the marking process entails. Greatorex, in her work on examiners' marking strategies, investigates how cognitive strategies change as examiners become more familiar with mark schemes and candidates' answers. In the third article on cognitive strategies, Crisp explores the judgement processes involved in A-level marking for both short answer questions and essays. The next four articles explore quality control of marking processes. Bell *et al*. outline new opportunities for quality control systems given the development of new technologies. Bramley offers a review of the terminology used to describe indicators of marker agreement and discusses some of the different statistics which are used in analyses. He goes on to consider issues involved in choosing an appropriate indicator and its associated statistic. In her study on double marking Rodeiro evaluates the agreement between marks from double marking models and discusses the advantages and disadvantages of blind and non-blind marking. The fourth article on quality assurance from Raikes and Massey focuses on the extent to which different examiners agree at item level and how far this agreement varies according to the nature of the item. This article contributes to the debate on the way in which new item level data, available due to advances in technology, could and should be used in future quality assurance procedures. In the final article Watts discusses the importance of fostering communities of practice amongst examiners in the context of new and developing technological systems and procedures.

**Sylvia Green** *Director of Research*

# The 'Marking Expertise' projects: Empirical investigations of some popular assumptions

**Dr Irenka Suto and Rita Nadas** Research Division

## 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 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 OCR's June 2005 examination papers. Our next two linked studies, which comprise *Marking Expertise Project 2*, are currently underway and involve both CIE and OCR examinations. 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 can be grouped according to which of these two routes they contribute. 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, 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 (Suto and Nadas, *in submission*). 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 maths, 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.
- 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.
- 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 were graduates. 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.

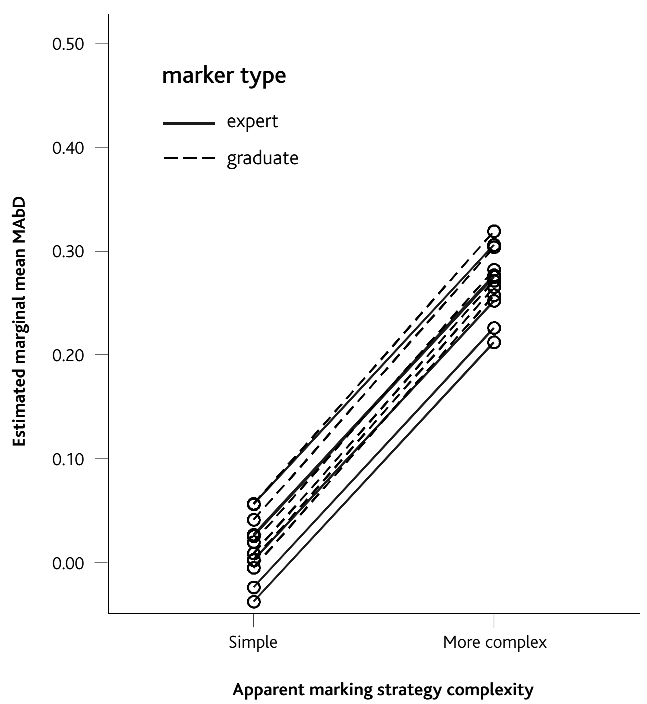



**Figure 1 : Graph showing estimated marginal mean MAbD values for individual physics markers (experts and graduates) for questions with different apparent marking strategy complexities**

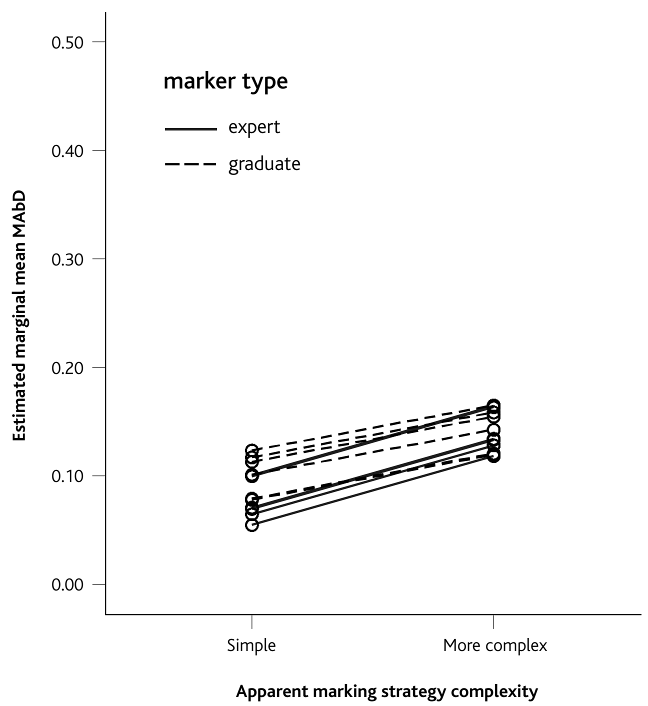



**Figure 2 : Graph showing estimated marginal mean MAbD values for individual maths markers (experts and graduates) for questions with different apparent marking strategy complexities**

### 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. 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 (PE) who participated in Study 1. These methods enabled the Principal Examiners to identify ways in which questions differed from one another, and thereby formulate distinct question features or 'constructs'. The Principal Examiners 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. 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.

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 Conference of International Association for Educational Assessment, Singapore.

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

Raikes N. (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. (*in submission*). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers.

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

# Did examiners' marking strategies change as they marked more scripts?

**Dr Jackie Greatorex** Research Division

## Introduction

Previously (Suto and Greatorex, *in press*) predicted that examiners might begin marking a question using particular cognitive strategies but later in the marking session they might use different cognitive strategies. My article describes a study designed to test this prediction. Changes in strategy usage might occur when examiners are more familiar with the mark scheme and candidates' answers. It is important to know whether examiners change their marking strategies because marking strategy usage might relate to the reliability and validity of marking. After all, Pinot de Moira *et al*. (2002) found varying degrees of inter- and intra-examiner reliability of marking at different times during the GCE A-level marking session of English. However, this is only one of many factors that can affect the reliability of marking.

There has been little research about the cognitive processes used to mark GCSEs, GCE A-levels and International GCSEs (IGCSE). To address this, Cambridge Assessment began a series of linked research projects. In one project examiners provided verbal protocols whilst marking GCSE Business Studies and GCSE Mathematics (Suto and Greatorex, *in press*). The researchers also conducted post-marking interviews with the examiners. The transcripts from the verbal protocols were analysed. From the analysis Suto and Greatorex (*in press*) reported five different cognitive strategies which examiners used to mark GCSEs. These were 'matching', 'scanning', 'evaluating', 'scrutinising' and 'no response'. Suto and Greatorex (*in press*) give a more detailed description of the strategies. Suto and Greatorex (2006) and Appendix 1 (p.11) give a concise description of the strategies. As this was an initial exploratory study the research studied the point in the marking process when examiners were familiar with the mark scheme, had marked a number of scripts and had experienced two co-ordination exercises. Subsequently, Greatorex and Suto (2006) undertook a further study of the cognitive strategies. One of our findings was that all of the five cognitive strategies were used to mark A-level Physics. Another of our findings was that there was no evidence of striking differences in the cognitive marking strategies used by examiners who were new to marking and by more experienced examiners.

The research about cognitive marking strategies drew from a psychological theoretical approach of dual processing – described in greater detail in Suto and Greatorex *(in press)*. This differentiates between two simultaneously active systems of cognitive processing. 'System 1' thought processes are automatic, quick, associative and intuitive. In contrast, 'system 2' thought processes are slow, effortful and reflective (Kahneman and Frederick, 2002; Stanovich and West, 2002). The different strategies entail using different processing systems (Suto and Greatorex, *in press*; Suto and Greatorex, 2006). 'Matching' and 'no response' entail simple system 1 type judgements. 'Scanning' utilises system 1 and/or system 2 type judgements. The 'evaluating' and 'scrutinising' strategies involve complex and reflective judgements (system 2 type judgements).

Kahneman and Frederick (2002) argue that as a person develops expertise and familiarity with a particular activity, cognitive operations might migrate from system 2 to system 1. This view describes how initially chess players have to think about the patterns on the board and what move to make, but how after much practice the players can recognise patterns more quickly and automatically make the appropriate moves. From these already established theories Suto and Greatorex (*in press*) predicted that examiners might begin marking a question using particular cognitive strategies but that later the examiners might use different cognitive strategies. For example, it is likely that examiners will use more 'matching' and 'scanning' when they are more familiar with the mark scheme and candidates' responses. Additionally, it is likely that examiners will use less 'evaluating' and 'scrutinising' when they are familiar with the mark scheme and candidates' responses. The present study was designed with the intention of investigating this prediction.

My research is an exploratory study dealing with small numbers of examiners. It involved five live[1] IGCSE examinations – Mathematics, Biology, French, Business Studies and English as a Second Language. The IGCSEs were taken by candidates in the autumn term of 2005. For each IGCSE candidates take a small number of assessments. The question papers used in this research included only one paper from each subject.

Some Biology questions required numerical skills, some required a short constructed prose response, some questions required graph work, another question required drawing a biological diagram. The Business Studies questions generally provided some information about a business situation and then asked for a short constructed written response. The notable exception was Q1aiii which involved each candidate drawing a graph. The English as a Second Language examination was a listening paper. The candidates were asked to listen to some spoken English and then give their responses to all of the questions. Some of the questions required short constructed prose responses and others true/false responses. The French examination contained some multiple choice questions, other questions required true/false responses and some further questions required a short constructed prose response. In the Mathematics examination some questions required stages of working and some included the use of diagrams. It was intended that these examinations would provide a good cross section of questions and mark schemes.

For these particular IGCSEs the Principal Examiners (PEs) wrote the question papers and led the marking. In larger examining teams the PEs ensured that the Team Leaders (TLs) were marking to the same standard as the Principal Examiner. The Team Leaders ensured the quality of the marking of the Assistant Examiners. In smaller examining teams there were no Team Leaders and the Principal Examiners monitored the quality of the Assistant Examiners' marking. Assistant Examiners initially marked a small number of scripts. The examiners then gathered at a co-ordination meeting and were briefed on how to apply the mark scheme. During the meeting examiners did some practice marking, and discussed some candidates' responses as well as discussing how to apply the mark scheme. By the end of each meeting a mark scheme was finalised. Subsequently, the Assistant Examiners each marked a *co-ordination sample* of scripts from their individual allocation of scripts. The co-ordination samples were then reviewed by a senior examiner to check the marking and to ensure that the Assistant Examiner could proceed with more marking. Later in the marking session two batches of marked scripts from each examiner's allocation were checked by a senior examiner. The first (batch 1) was compiled after about 40% of the Assistant Examiner's marking was complete and the second (batch 2) was compiled from the remainder of their marking. Both the total score the senior examiner gave to each script and the total score the Assistant Examiner gave to each script were recorded on a form which was returned to CIE. If their marking was not sufficiently similar then action was taken.

I reported elsewhere that telephone interviews were undertaken with examiners from Mathematics and Biology (Greatorex, 2006). The purpose of the interviews was to establish which cognitive strategies were used during marking. I found that the cognitive strategies used by examiners in other GCSE and UK A-level subjects were being used to mark IGCSE Mathematics and Biology in the winter 2005 session. So it was hoped that the strategies were relevant to the French, English as a Second Language and Business Studies examinations described above. A questionnaire was used to study any patterns of changes in marking strategies in a wider group of examiners and subjects.

## Method

### Questionnaire development

A questionnaire was developed which referred to the different parts of the marking session described above. The questionnaire was piloted with a Business Studies examiner from a GCSE syllabus. The pilot indicated that the questionnaire was valid and practical. But the pilot was not sufficient to establish how well each questionnaire question worked from a psychometric viewpoint. Furthermore, Awarding Body staff with experience in writing and administering questionnaires to examiners, candidates and centres reviewed the questionnaire. The questionnaires asked about different occasions in the marking session:

- before the co-ordination meeting
- during the co-ordination sample
- during batch 1
- after batch 1

The questionnaire was adjusted slightly for each subject. See Appendix 2 (p.12) for an example of the questionnaire.

The questionnaire focussed on a selection of examination questions (see Table 1) to ensure that it was manageable and covered the range of question types. I chose these examination questions because I thought that at least one question from each subject entailed examiners using system 1 thought processes and at least one further question from the same subject involved examiners using system 2 thought processes.

**Table 1 : The examination questions included in the questionnaire**

| *Examination* | *Examination Question* |
|---|---|
| Biology | 1aiv, 1ci, 3a |
| Business Studies | 1aii, 3ai, 4 |
| English as a Second Language | 1, 6, 7 |
| French | 1, 26, 31 |
| Mathematics | 1, 11, 21b |

### Administration

The questionnaire was administered in January 2006 when the marking was complete. All examiners received a definition of each of the five strategies (see Appendix 1) as well as subject specific materials (the questionnaire, the question paper, and the mark scheme). The participants were asked to read all the materials provided before answering the questionnaire.

### Participants

All Principal Examiners (n=5), Team Leaders (n=5) and Assistant Examiners (n=59) who marked in the November and December 2005 marking session were sent the materials. Table 2 gives the number of examiners who marked in the session. The number of research participants that responded to the questionnaire is given in brackets. Note that Table 2 gives figures regarding *all* examiners; no distinction is made between the senior examiners and the Assistant Examiners. Some of the participants had marked these examinations a number of times before and others were new to marking the examination.

**Table 2 : Summary of the number of examiners who marked and responded to the questionnaire**

| *Examination* | *Total number of examiners that marked (total number of examiners that responded)* |
|---|---|
| Biology | 10 (7) |
| Business Studies | 26 (19) |
| English as a Second Language | 7 (6) |
| French | 6 (5) |
| Mathematics | 20 (13) |

## Results

This section reports on the responses that the examiners gave to part of the questionnaire. I present extracts from the question papers and mark schemes to facilitate the readers' understanding of the results (see below). There is also a graph summarising some of the examiners' responses to the questionnaire. A commentary for each graph is provided below to highlight (1) the relative proportion with which each strategy was used when all of the marking session was considered; and (2) any differences in ratings (strategy usage) between consecutive occasions.

In this analysis the term 'strategy usage' is used as a shorthand phrase for the self-reported perceived strategy usage indicated by the examiners' ratings. A change in strategy usage (ratings) of 33% or more for one strategy is described as a 'considerable' difference (change). A change in strategy usage of about 20% or more is described as a 'noticeable' difference (change). These percentages and definitions are somewhat arbitrary. Differences were calculated by subtracting the percentage of responses (rating) of 'never' for 'matching' from the percentage of responses of 'never' for 'matching' for a consecutive

1. 'Live' means that the examinations were real and not taken in an experimental setting.

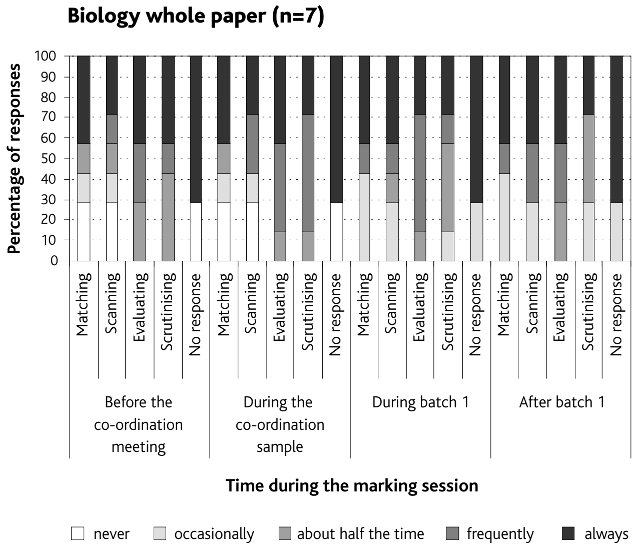



**Figure 1 : Graph to show the percentage of ratings for Biology examiners**

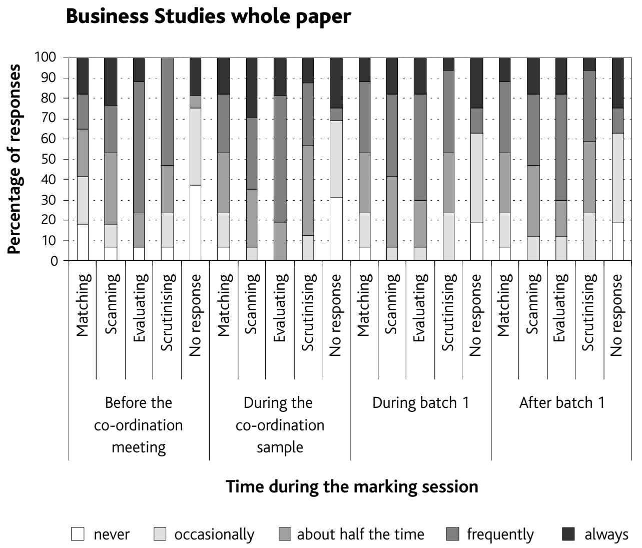



**Figure 2 : Graph to show the percentage of ratings for Business Studies examiners**

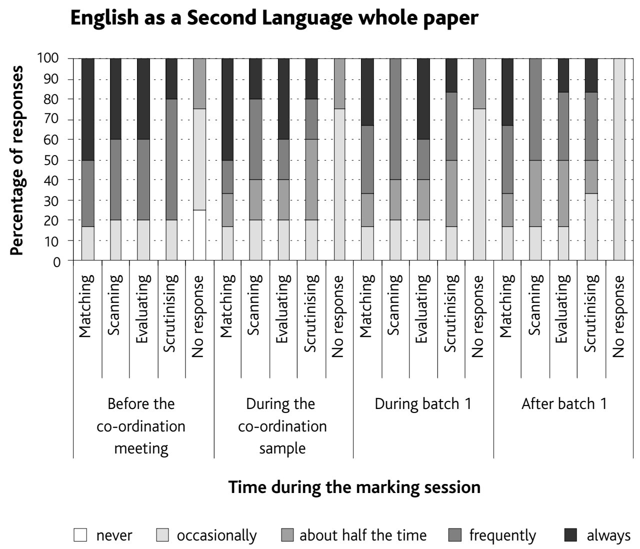



**Figure 3 : Graph to show the percentage of ratings for English as a Second Language examiners**

occasion. This was repeated for each response category, strategy, occasion and questionnaire item. The easiest way to make inferences from the information in the following graphs is to bear in mind that the darker a bar the more that strategy was used on that occasion.

For the sake of brevity the *Results* section only presents some of the key findings. I chose these particular key findings to illustrate the points made in the *Conclusion and Discussion*. For a more detailed report of the findings see Greatorex (2006).

## Biology whole examination

The data presented in Figure 1 illustrate that for the whole Biology examination the 'evaluating' strategy had the largest proportion of 'always' and 'frequently' ratings, followed by 'no response', 'scanning', 'scrutinising' and then 'matching'. Regarding differences in strategy usage on consecutive occasions, there was a large difference in the ratings on 'scrutinising', from which it can be inferred that more 'scrutinising' was being used during the co-ordination sample than before the co-ordination meeting or during batch 1. There were some noticeable differences in the ratings about 'matching', 'scanning' and 'no response'; these differences imply that these strategies were used more often during batch 1 than during the co-ordination sample.

## Business Studies whole examination

The data presented in Figure 2 illustrate that for the whole Business Studies examination the 'evaluating' strategy had the largest proportion of 'always' and 'frequently' ratings. The strategy with the next largest proportion of these ratings was 'scanning', followed by 'scrutinising' and then 'matching'. 'No response' was the strategy with the smallest proportion of 'always' and 'frequently' ratings. Regarding differences in ratings between consecutive occasions there were no considerable differences. There was a noticeable difference in the ratings about 'scrutinising', which implies that the 'scrutinising' strategy was used more in the co-ordination sample than before the co-ordination meeting.

## English as a Second Language whole examination

The data presented in Figure 3 indicate that for the whole English as a Second Language examination the 'matching' strategy had a larger proportion of 'always' and 'frequently' ratings. The 'scanning' and 'evaluating' strategies each had slightly smaller proportions of these ratings and the 'scrutinising' strategy had an even smaller proportion. The 'no response' strategy had zero 'always' and 'frequently' ratings. Regarding differences in ratings between consecutive occasions there was a considerable difference in the ratings about the 'scrutinising' strategy. The difference in ratings implied that 'scrutinising' was used more before the co-ordination meeting than during the co-ordination sample. There were some noticeable differences in ratings for the 'no response', 'evaluating' and 'scanning' strategies. From the differences it can be inferred that:

- the 'no response' strategy was used more during the co-ordination sample than before the co-ordination meeting;
- the 'evaluating' strategy was used more during batch 1 than afterwards, and more before the co-ordination meeting than during the co-ordination sample;
- the 'scanning' strategy was used more before the co-ordination meeting than during the co-ordination sample, and more during the co-ordination sample than during batch 1.

## French whole examination

The data presented in Figure 4 illustrate that for the whole paper the 'evaluating' strategy had the larger proportion of 'always' and 'frequently' ratings, followed by 'scanning', 'matching' and then 'scrutinising'. 'No response' had the smallest proportion of these ratings. There was a considerable difference in ratings on consecutive occasions. From this difference it can be inferred that the 'scrutinising' strategy was used more during the co-ordination sample than during batch 1. There was also a noticeable difference in ratings on consecutive occasions. This difference implied that the 'matching' strategy was used more during the co-ordination sample than before the co-ordination meeting.

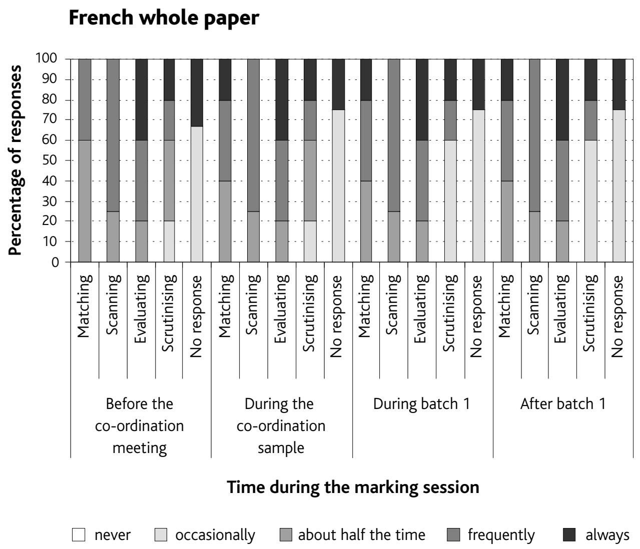
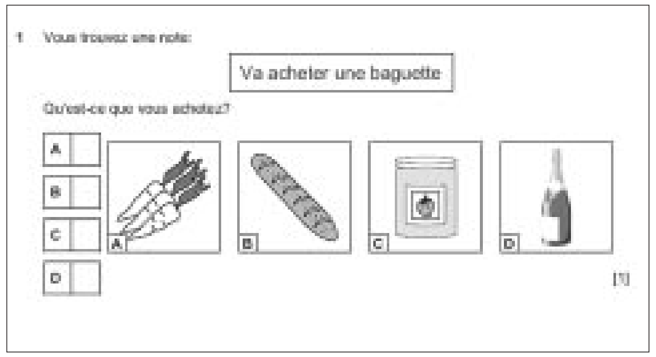



**Figure 4: Graph to show the percentage of ratings for French examiners**


1 Vous trouvez une note:

Va acheter une baguette

Qu'est-ce que vous achetez?

A B C D

[1]

**French Question Paper extract**

| | |
|---|---|
| 1 B | 1 Mark per item |

**Mark Scheme extract**

The data presented in Figure 5 indicate that for question 1 the 'matching' strategy was the only strategy with 'always' and 'frequently' ratings. Regarding differences in ratings between consecutive occasions there was one considerable difference, from which it can be inferred that the 'scrutinising' strategy was used less during the co-ordination sample than before the co-ordination meeting. There were also some noticeable differences in ratings which implied that the 'scrutinising' strategy was used more during batch 1 than during the co-ordination sample and that the 'matching' strategy was used more during the co-ordination sample and before the co-ordination meeting than later in the marking session.

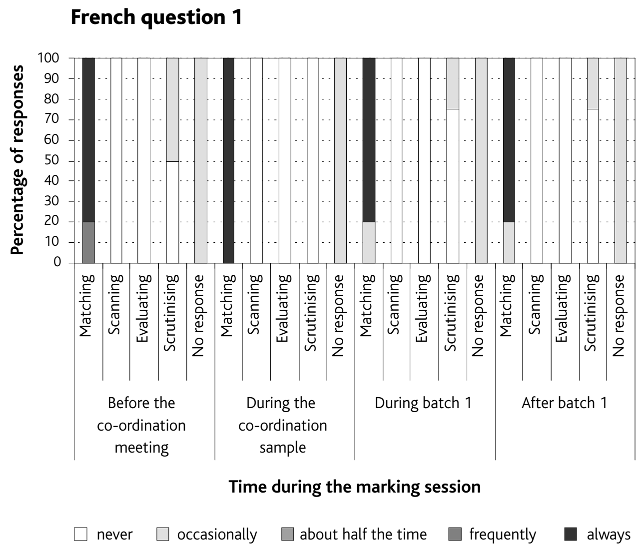



**Figure 5 : Graph to show the percentage of ratings for French examiners**

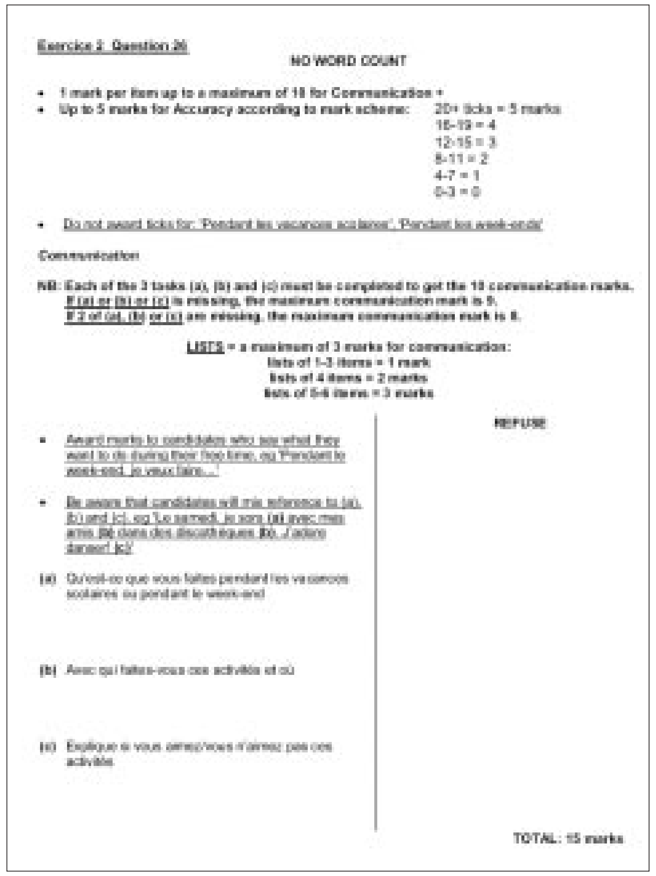

26 Le Temps libre

Écrivez environ 80-90 mots **en français** pour dire:

(a) Qu'est-ce que vous faites pendant les vacances scolaires ou pendant le week-end?
(b) Avec qui faites-vous ces activités et où?
(c) Expliquez si vous aimez/vous n'aimez pas ces activités.

[Total : 15]

**French Question Paper extract**


Exercice 2 Question 26

NO WORD COUNT

- 1 mark per item up to a maximum of 10 for Communication +
- Up to 5 marks for Accuracy according to mark scheme: 20+ ticks = 5 marks
  16–19 = 4
  12–15 = 3
  8–11 = 2
  4–7 = 1
  0–3 = 0
- Do not award ticks for: 'Pendant les vacances scolaires', 'Pendant les week-ends'

Communication

NB: Each of the 3 tasks (a), (b) and (c) must be completed to get the 10 communication marks.
If (a) or (b) or (c) is missing, the maximum communication mark is [illegible].
If 2 of (a), (b) or (c) are missing, the maximum communication mark is [illegible].

LISTS = a maximum of 3 marks for communication:
lists of 1–3 items = 1 mark
lists of 4 items = 2 marks
lists of 5-6 items = 3 marks

REFUSE

- Award marks to candidates who say what they want to do during their free time, eg 'Pendant le week-end, je veux faire...'
- Be aware that candidates will mix reference to (a), (b) and (c), eg 'Le samedi, je sors (a) avec mes amis (b) dans des discothèques (b). J'adore danser! (c)'

(a) Qu'est-ce que vous faites pendant les vacances scolaires ou pendant le week-end

(b) Avec qui faites-vous ces activités et où

(c) Expliquez si vous aimez/vous n'aimez pas ces activités

TOTAL: 15 marks

**Mark Scheme extract**

The data presented in Figure 6 indicate that for question 26 the 'evaluating' strategy had a larger proportion of 'always' and 'frequently' ratings. The 'scanning' strategy had the next largest proportion, followed by the 'scrutinising' and 'matching' strategies. The 'no response' strategy had zero 'always' and 'frequently' ratings. Regarding differences in ratings on consecutive occasions there were two considerable differences which implied that the 'matching' strategy was used less and the 'scanning' was used more during batch 1 than during the co-ordination sample. There were a number of noticeable differences in the ratings. These differences implied that the 'evaluating' strategy was used more and the 'scrutinising' strategy was used less before the co-ordination meeting than during the co-ordination sample. From the differences it can also be inferred that the 'evaluating' and 'scrutinising' strategies were used more and the 'no response' strategy used less during the co-ordination sample than during batch 1.

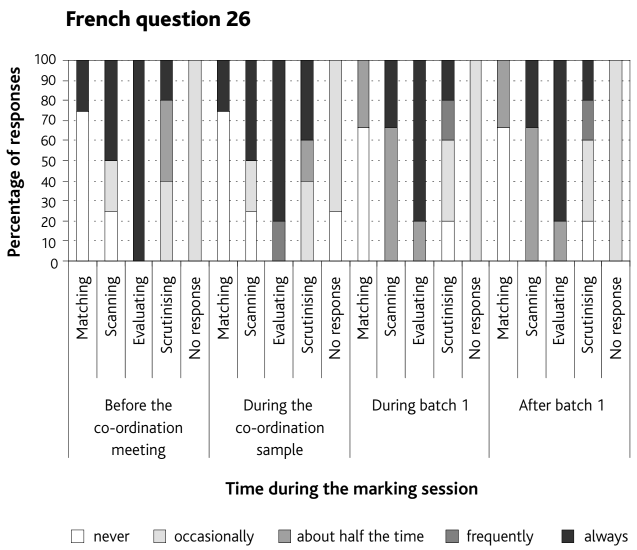



**Figure 6 : Graph to show the percentage of ratings for French examiners**

## Mathematics whole examination

The data presented in Figure 7 illustrate that for the whole Mathematics examination the 'matching' strategy had the larger proportion of 'always' and 'frequently' ratings. The other strategies, 'scrutinising', 'no response', 'evaluating' and 'scanning', are given in descending order of the relative size of the proportion of 'always' and 'frequently' ratings. There were no considerable differences in ratings on consecutive occasions. However, it can be inferred from inspecting Figure 7 that the 'scanning', 'evaluating' and 'scrutinising' strategies were all used less during and after batch 1 in comparison to the beginning of the marking session.

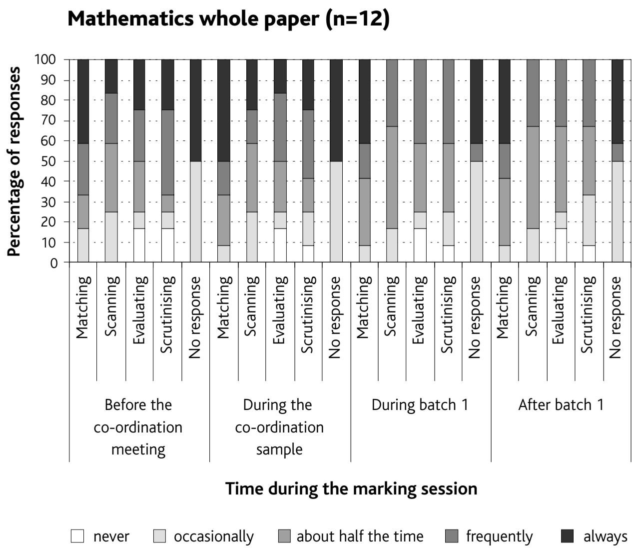



**Figure 7 : Graph to show the percentage of ratings for Mathematics examiners**

## Conclusion and Discussion

The research was limited by some factors.

First, as with many self report methods the retrospective questionnaire ratings of how often the examiners estimated they used a particular strategy are limited. After all, some examiners will be better than others at making estimates. Additionally, the strategy usage depends on what the candidate has written, as well as the personal preferences of examiners, along with examiners' various marking or teaching experiences, and examiners' ability to choose appropriate strategies.

Secondly, the senior examiners' ratings might have been different to the Assistant Examiners' ratings. The graphs (above) mask any differences between the ratings from the two contrasting groups of examiners. The senior examiners were included in the graphs with the Assistant Examiners as the number of examiners was so low (as some of the examining teams were small).

Thirdly, for the purpose of gaining background information about the examination process I attended the French co-ordination meeting.



During this meeting there was some discussion about marking strategies which might have given the examiners who marked French more background information to complete the questionnaire. All examiners in all subjects were provided with a description of the strategies for the purposes of completing the questionnaire.

Fourthly, it was not clear to what extent the research results about changes in proportionate strategy usage can be generalised beyond the distinctive IGCSE examination context.

'Scanning' utilises system 1 and/or system 2 type judgements. Therefore, previous literature could not be used to make predictions about how often the 'scanning' strategy might be used at different stages in the marking session. Additionally, differences in ratings about 'scanning' might imply a change from mostly system 2 to primarily system 1 processing. Alternatively, the same differences might imply a change from mostly system 1 to primarily system 2 processing. Therefore, I cannot make inferences about any changes in the proportion of system 1 and system 2 type judgements that were used. For more details about the scanning strategy see Suto and Greatorex (2006, *in press*).

In the following section examiners are treated as a group; I am not considering differences between individual examiners.

Any changes in the number of times examiners reported using the 'no response' strategy depended on the content of the scripts marked on that occasion.

In my research examiners used all or most of the strategies, for each question, when the whole marking session was considered (e.g. see Figure 6). However, as expected, there were some questions for which the ratings implied that a particular strategy was overwhelmingly used, for example, 'matching' for question 1 in the French examination (Figure 5). My research findings are in line with those of Suto and Greatorex (*in submission*) who found that for some individual Business Studies and Mathematics questions one strategy was overwhelmingly used, but that for other questions a combination of strategies were employed. Now we have further evidence that strategy usage varies for individual questions.

In previous research we found that all the marking strategies were used to a greater or lesser extent to mark GCSE Business Studies and GCSE Mathematics, as well as Physics A-level (Greatorex and Suto, 2006). In my research we can infer from the ratings that all strategies were used to mark the Biology, Business Studies, English as a Second Language and Mathematics examinations. The ratings also imply that there was some variability in the extent to which each strategy was used to mark each IGCSE examination; there was no strategy that was used overwhelmingly often to mark a particular examination (e.g. Figure 1, Figure 2, Figure 3, and Figure 7). My research highlights that the strategies reported by Suto and Greatorex (2006, *in press*) are used to a greater or lesser extent to mark a wider variety of examinations and qualifications than was previously evidenced.

The research was designed to test whether examiners begin marking a question using particular cognitive strategies but later they might use different cognitive strategies. Kahneman and Frederick (2002) argue that as a person develops expertise and familiarity with a particular activity, cognitive operations can migrate from system 2 to system 1. As already mentioned, the 'evaluating' and 'scrutinising' strategies involve complex and reflective judgements (system 2 type judgements). Therefore, Suto and Greatorex (*in press*) predicted that examiners might use less 'evaluating' and 'scrutinising' when they are familiar with the question paper, mark scheme and candidates' responses. The 'matching' strategy entails simple system 1 type judgements. Therefore, Suto and Greatorex (*in press*) also predicted that examiners might use more 'matching' when they are familiar with the question paper, mark scheme and candidates' responses. In my research there were not many *considerable* differences in ratings between consecutive occasions, so there were not as many changes in strategy usage as we had predicted. However, when there were *considerable* differences these were mostly in the direction we predicted. For example, Figure 4 illustrates a considerable decline in the use of 'scrutinising' from the co-ordination sample to batch 1. To see this difference the reader needs to study the graph closely. Please note that 60% of the bar in Figure 4, referring to using scrutinising during the co-ordination sample, is made up of 'about half the time' and 'occasionally' ratings. But 60% of the bar in Figure 4, referring to using scrutinising during batch 1, constitutes 'occasionally' ratings. This is one of the considerable differences I found in strategy usage.

Many research questions were not addressed by my research or previously published studies. For instance, (1) what cognitive strategies are used to mark other subjects and groups of questions, particularly those with longer questions or even Art or aesthetic subjects?, and (2) does examiners' ability to choose appropriate marking strategies vary? However, my research highlights that sometimes examiners' marking strategies changed as the examiners marked more scripts.

APPENDIX 1

## Marking Strategies Reference Sheet (updated)

In previous research a colleague and I (Suto and Greatorex, *in press*, 2006) reported that there are a number of strategies that examiners use to mark. In the research examiners were asked to 'think aloud' whilst they were marking. The strategies are described below and are illustrated with an example.

### Matching

*Matching* can be used when the answer to a question part is a visually recognisable item or pattern, for example, a letter, word, number, part of a diagram, short sequence of words or letters. The examiner looks at a short answer line or other pre-determined location and compares the candidate's response with the correct answer (either held in the working memory or recollected using the mark scheme), making a simple judgement about whether they match.

| *Question paper extract* | *Mark scheme extract* |
|---|---|
| State whether the following statement is true or false<br>The Euro is another name for the European Union_______[1] | False (1) |

To mark this question examiners were looking at the short answer line and comparing the mark scheme answer 'false' to the candidate's answer.

### Scanning

*Scanning* occurs when an examiner scans the whole of the space in the script allocated to a question part, in order to identify whether a particular detail in the mark scheme is present or absent. This detail might be a letter, word, part of a diagram or similar. When the scanned-for detail is simple (e.g. a single number or letter), pattern recognition occurs. When the scanned-for detail requires more meaningful or semantic processing, for example, a stage of mathematical working, an additional marking strategy thought process may need to be used.

For one question, when the examiners predominantly used scanning, they searched the candidate's answer in the whole of the answer space for stages of working, for example, '2.29-0.021'.

### Evaluating

When *evaluating*, an examiner attends to either all or part of the space dedicated to an item. He or she processes the information semantically, considering the candidate's answer for structure, clarity, factual accuracy and logic or other characteristics given in the mark scheme. Sometimes a single judgement about the mark value for a particular answer is made at the end of evaluating a response. At other times, one or more interim judgements of the appropriate mark value for the candidate's answer are made during the evaluation process.

In one question candidates were given detailed information about a company and its situation along with four options A, B, C and D for what the company could do next. Candidates were asked to discuss which of options A, B, C or D would be best for the company. There were 8 marks available. To mark the question examiners used the evaluating strategy. Whilst one examiner was thinking aloud, they said first that as they were reading the answer they saw that a candidate had identified two options, each of which the examiner judged the candidate gave one sided support. Secondly, the examiner found that the candidate identified a third option which the examiner judged the candidate had analysed. Thirdly, the examiner decided that the candidate made some general comments but did not make an overall conclusion. The examiner gave the candidate the appropriate credit.

### Scrutinising

*Scrutinising* follows on from, or is used in conjunction with, one of the other strategies, and is used only when a candidate's answer is unexpected or incorrect. The examiner tries to identify where the problem lies and whether a valid alternative to the mark scheme solution has been given. To do this, he or she evaluates multiple aspects of the candidate's response with the overarching aim of reconstructing the candidate's line of reasoning or working out what the candidate was trying to do. The examiner may have to deal with a lot of uncertainty and re-read the candidate's response several times.

### No response

The *no response* strategy is used when a candidate has written nothing in the answer space allocated to the question part. The examiner looks at the space once or more to confirm this; he or she can award 0 marks for that item.

APPENDIX 2

## Questionnaire – Process of marking – French

INSTRUCTIONS

The 'marking strategies reference sheet', question paper and mark scheme are provided for reference. You will need to read the 'marking strategies reference sheet' before answering this questionnaire.

Please indicate for each examination question how often you use each strategy when marking for each stage of the marking process. Please write
"0" to indicate "never"
"1" to indicate "occasionally"
"2" to indicate "about half the time"
"3" to indicate "frequently"
"4" to indicate "always"

| | | *Before the co-ordination meeting* | *During the co-ordination sample* | *during batch 1* | *after batch 1* |
|---|---|---|---|---|---|
| **Question 1** | 'matching' | | | | |
| | 'scanning' | | | | |
| | 'evaluating' | | | | |
| | 'scrutinising' | | | | |
| | 'no response' | | | | |
| **Question 26** | 'matching' | | | | |
| | 'scanning' | | | | |
| | 'evaluating' | | | | |
| | 'scrutinising' | | | | |
| | 'no response' | | | | |
| **Question 31** | 'matching' | | | | |
| | 'scanning' | | | | |
| | 'evaluating' | | | | |
| | 'scrutinising' | | | | |
| | 'no response' | | | | |

Please estimate for the whole examination how often you use each strategy when marking for each stage of the marking process. Pease make an overall estimate rather than making judgements for every question and then estimating totals.

| | | *Before the co-ordination meeting* | *During the co-ordination sample* | *during batch 1* | *after batch 1* |
|---|---|---|---|---|---|
| **Whole examination paper** | 'matching' | | | | |
| | 'scanning' | | | | |
| | 'evaluating' | | | | |
| | 'scrutinising' | | | | |
| | 'no response' | | | | |



**Acknowledgements**

Jane Fidler has worked on some of the data management and presentation of images, graphs, figures and tables in this article. Rita Nadas has also worked on some of the data management in this study.

**References**

Greatorex, J. (2006). *Do examiners' approaches to marking change between when they first begin marking and when they have marked many scripts?* A paper presented at the British Educational Research Association Annual Conference, September 2006, University of Warwick.

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 of Educational Assessment conference, May 2006, Singapore.

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

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

Stanovich, K. & West, R. (2002). Individual differences in reasoning. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press.

Suto, W. M. I. & Greatorex, J. (*in submission*). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations.

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. & Greatorex, J. (2006). A cognitive psychological exploration of the GCSE marking process. *Research Matters: A Cambridge Assessment Publication*, **2**, 7–10.

PSYCHOLOGY OF ASSESSMENT

# Researching the judgement processes involved in A-level marking

**Victoria Crisp** Research Division

## Introduction

The marking of examination scripts by examiners is the fundamental basis of the assessment process in many assessment systems. Despite this, there has been relatively little work to investigate the process of marking at a cognitive and socially-framed level. Vaughan (1991) and others have commented on the importance of investigating the process and decision-making behaviour through which examiners make their evaluations. According to Milanovic, Saville and Shuhong (1996), the lack of understanding about the decision-making process makes it hard to train examiners to make valid and reliable judgements. A decade later their view is still accurate. Improved understanding of the judgement processes underlying current assessment systems would also leave us better prepared to anticipate the likely effects of various innovations in examining systems such as moves to on-screen marking.

The research summarised here started by reviewing the relevant literature in the areas of cognitive judgement, theories of reading comprehension, social theories of communities and research specifically investigating the decision-making and judgements involved in marking. Notable amongst the latter are the works of Vaughan (1991), Pula and Huot (1993) and Huot (1993) in the context of assessing writing, Milanovic, Saville and Shuhong (1996), Cumming (1990) and Lumley (2002) in the context of English as a second language, Sanderson (2001) with regard to marking A-level sociology and law essays and Greatorex and Suto (2006) in the context of short answer questions in maths and business studies GCSE papers. Few studies have researched the marking of disciplines other than English writing and none have considered the processes involved in marking short answer questions and essays within the same domain. This research was designed and undertaken to address this gap in our understanding of examiners' judgements and attempted to draw on a wider range of relevant theoretical areas than have been used in most previous studies.

## Method

An AS level and an A2 level geography exam paper were selected. The AS level exam required students to provide short to medium length responses whilst the A2 unit involved writing two essays from a choice. Six experienced examiners who usually mark at least one of the two papers participated in the research. Five of the examiners were usually only involved in the marking of one of the papers but most had experience of teaching both units and would be eligible to mark the other.

Examiners marked fifty scripts from each exam at home with the marking of the first ten scripts for each reviewed by the relevant Principal Examiner. This reflected normal marking procedures as far as possible but the marking was not subject to the same degree of standardisation as live marking. Examiners later came to meetings individually where they marked four or five scripts in silence and four to six scripts whilst thinking aloud for each exam, and were also interviewed.

The scripts marked were photocopies of live scripts with marks and annotations removed. Examiners marked the same students' scripts, except that in the silent marking and think aloud marking, for each examiner one of the scripts in each batch was a clean copy of one of the scripts included in the main batch of home marking.

## Results

Analysis of the marks awarded during the home marking suggested that marking was broadly in line with live marking but that examiners tended towards severity in comparison. One examiner's marking of the AS unit was more severe than the others' and out of line with live marking and the same was the case for a different examiner's marking of the A2 unit.

The analysis of mark changes between home marking and silent marking at the meeting, and between home marking and marking whilst thinking aloud for the small number of repeated scripts suggested that thinking aloud affected the marks awarded very little, if at all. Thinking aloud seemed to result in slightly more consistent marking for short and medium length responses and slightly less consistent marking with essays, but these differences were small and could have occurred by chance. This helps to confirm that verbal protocol analysis is a valid research method in the investigation of the judgements involved in exam marking.

### Coding the verbal protocols

Transcripts of the verbal protocols were analysed to try to understand the processes involved in the marking. Drawing on the transcripts and the work of Sanderson (2001) and Milanovic *et al*. (1996) a detailed coding frame was developed to reflect the specific qualities of student work noticed by markers and marker behaviours and reactions. The codes were grouped into the categories of:

- 'reading and understanding' (codes relating to reading and making sense of responses);
- 'evaluates' (codes relating to evaluating a response or part of a response);
- 'language' (codes relating to the student's use of language);
- 'personal response' (affective and personal reactions to student work);
- 'social perception' (social reactions such as making assumptions about candidates, talking to or about candidates, comments about teaching);
- 'task realisation' (codes relating to whether a student has met the demands of the task such as length of response, addressing/understanding question);
- 'mark' (codes relating to assessment objectives and quantifying judgements).

These categories are described in a little more detail below with short quotes from the verbal protocols included to exemplify the behaviours/reactions being described where this is helpful.

### Reading and understanding

Not unexpectedly, reading and interpretation behaviours were frequent in the verbal protocols, perhaps emphasising the sometimes over-looked importance of reading and making sense of a student's intended meaning as a prerequisite to accurate and valid marking. Codes in this category included, among others, obvious reading behaviours, summarising or paraphrasing all or part of a response and seeking or scanning for something in particular in the student's work (e.g. '*really we are looking for two regions well described and explained to illustrate that unevenness*').

### Evaluating

Also frequently apparent in the verbal protocols (and not unexpected) were behaviours relating to evaluating the text being read. Clearly positive and negative evaluations (e.g. '*so that's a good evaluation point', 'no she hasn't got the correct definition of site, she is confusing it'*) were particularly frequent whilst other behaviours such as weighing up the quality of part of a response and making comparisons with other responses were also apparent.

### Language

For both exam papers, all examiners made some comments about the quality of the language used by students. Some examiners also commented on the orthography (handwriting, legibility and general presentation) of student work, particularly with the essay paper (e.g. '*bit of a difficulty to read this towards the end, he has gone into scribbly mode*'). Comments on language and orthography were often negative.

### Personal response

This category was created to accommodate the affective (i.e. emotional) reactions of some examiners to student work that sometimes occurred and other personal comments or responses. Reactions in this category included positive or negative affect (e.g. *'I quite like that', 'London [groan] my heart drops'*), laughter and frustration or disappointment. All examiners showed one or more of these reactions at some point but behaviours in this category were generally fairly infrequent except in the case of one examiner.

### Social perception

Examiners sometimes displayed reactions associated with social perceptions of the imagined candidates. Markers sometimes made assumptions about the likely characteristics of the candidate (e.g. '*clearly not a very bright candidate*'), predicted further performance of the candidate (e.g. '*this is not going to be a better paper is it?*') and talked to or about the candidate, sometimes almost entering into a dialogue with the student via the script (e.g. '*so give us an example now of this*'). Comments about teaching were also grouped into the category. Social perception type behaviours were generally fairly infrequent and varied in frequency between examiners, perhaps reflecting the personalities of individual examiners.

### Task realisation

The comments coded in this category were about features of the responses required of students in order to successfully achieve the task set by a question and included comments on the length of responses, on material missing from a student's response (e.g. '*that doesn't really say why and it doesn't use map evidence*'), on the relevance of points made and on whether the candidate has understood and addressed the intended question.

### Mark

A number of different types of behaviours relating to quantifying evaluations and making a mark decision were observed. These included (among other behaviours) comments regarding the Assessment Objectives stated in the mark scheme (particularly for the A2 exam), initial indications of marks, reference to the mark scheme or to marking meetings or to the Principal Examiner's guidance and reflections on the total mark scored or on their own leniency or severity.

The following table shows a transcript extract from an examiner's marking of a short answer response along with the codes that were applied to this extract.

| *Transcript extract* | *Codes* |
|---|---|
| *Now we have got Mexico, Mexico city from rural areas, ok,* | Summarises/paraphrases positive evaluation |
| *increasing at a rate, mentions job opportunities, well explained there,* | Summarises/paraphrases positive evaluation |
| *a high standard, cramped housing, talking about what it is like in Mexico city rather than the whole country, (.)* | Summarises/paraphrases neutral evaluation |
| *shanty towns, now it's gone on to talk,* | Summarises/paraphrases |
| *most of it is irrelevant there,* | Negative evaluation and relevance |
| *but, let's have a look and see, explanation in only one area, [using mark scheme] (.)* | Reference to mark scheme |
| *so it's level 2 and is fairly general* | First indication of mark |
| *so I think we will give that 5* | Mark decision |
| *because it hasn't really explained much more than, not a lot about what it is like where they've come from, so really only explaining one area, southern* | Discussion/review of mark/re-assessment |

## Findings

**Did the frequencies of coded behaviours and reactions vary between the marking of different types of questions (short and medium length questions versus essays)?**

The frequencies of codes were compared between the exam papers in order to consider whether there were differences in the behaviours involved in marking short to medium length responses and marking essays. There was no significant difference in the average total number of codeable behaviours per script between the two exams but there were a number of differences in the frequencies of individual codes. Differences included greater frequencies of two codes relating to social perceptions (assumptions about characteristics of candidates, predicting further performance) with the essay paper than with the AS exam. In addition, there were more instances of comments about addressing the question and about orthography (handwriting, legibility, presentation) with the A2 exam and greater acknowledgement of missing material with the AS exam. There were also differences in the frequencies of 'mark' related codes with more frequent reference to assessment objectives in the A2 exam, and more frequent occurrence of other mark related codes such as 'first indication of mark', 'discussion/review of mark/re-assessment' with the AS unit due to the greater number of mark decisions that have to be made. Examiners more frequently reflected on the total mark when marking the essay paper than with the shorter answer paper.

These differences give us some insight into the areas in which there might be a greater risk of examiner bias for each type of exam paper. There is more potential for assumptions about candidates or predicting performance in advance of a full reading to cause bias with essays than with shorter questions. There may be more risk of poor handwriting causing bias with essays. In addition, examiners are more likely to focus on what is missing from shorter responses than with essays. This is not to say that there was clear evidence of examiner bias occurring in these areas or that these are significant areas of risk but that these may be areas of potential risk worth bearing in mind when planning examination specifications and in marker training.

**Did the frequencies of different types of behaviours and reactions vary between different examiners?**

Differences between examiners in the frequencies of occurrence of codes were found for 31 of the 42 codes. Despite the variations in the frequencies of occurrence of individual behaviours or reactions between examiners, it seems that in most instances these differences did not have a significant impact on the agreement of marks between markers and that different marking styles can be equally effective.

Detailed analysis of the behaviours evidenced in the verbal protocols of the two examiners (one with the AS exam and one with the A2 exam) who awarded total marks that were significantly different to those of the other examiners offered some tentative hypotheses about influences on reliability. For example, greater frequencies of first indications of marks and discussion of marks were associated with lower marker agreement for one examiner which might suggest that over-deliberating on marking decisions is not advantageous. Lower frequencies of obvious reading behaviours were associated with lower marker agreement for both examiners, as were lower frequencies of comparisons with other scripts/responses and lower frequencies of positive evaluations.

**Did the frequencies of coded behaviours and reactions vary between questions and/or between scripts?**

Differences in the frequencies of code occurrence between questions were found for around half of the codes and were often associated with one particular essay question on a popular topic. There were few differences between scripts in the frequencies of codes that were applied suggesting that marking behaviours for different students' scripts were similar and that the findings are likely to be generalisable to other students' scripts beyond the sample used in the research. It seems that the processes involved in marking are infrequently affected by features of the scripts.

**Which codes frequently occurred together?**

Considering the frequently co-occurring codes also provided some interesting findings. Evaluations were often associated with aspects of task realisation (e.g. missing material, addressing/understanding question) and with the assessment objectives. Additionally, evaluations (especially negative evaluations) were often associated with considerations of the marks to be awarded. Positive evaluations and negative evaluations often co-occurred reflecting instances where examiners considered the strengths and weaknesses of a response or part of a response (e.g. '*a vague comment about the relief of the area*').

## Towards a model of the marking process

Analysis of the sequences of the coded behaviours apparent in the verbalisations allowed a tentative model of the marking process to be constructed. The model conceptualises three main phases and less frequently occurring 'Prologue' and 'Epilogue' phases before reading begins and after mark decisions have been made. The model attempts to

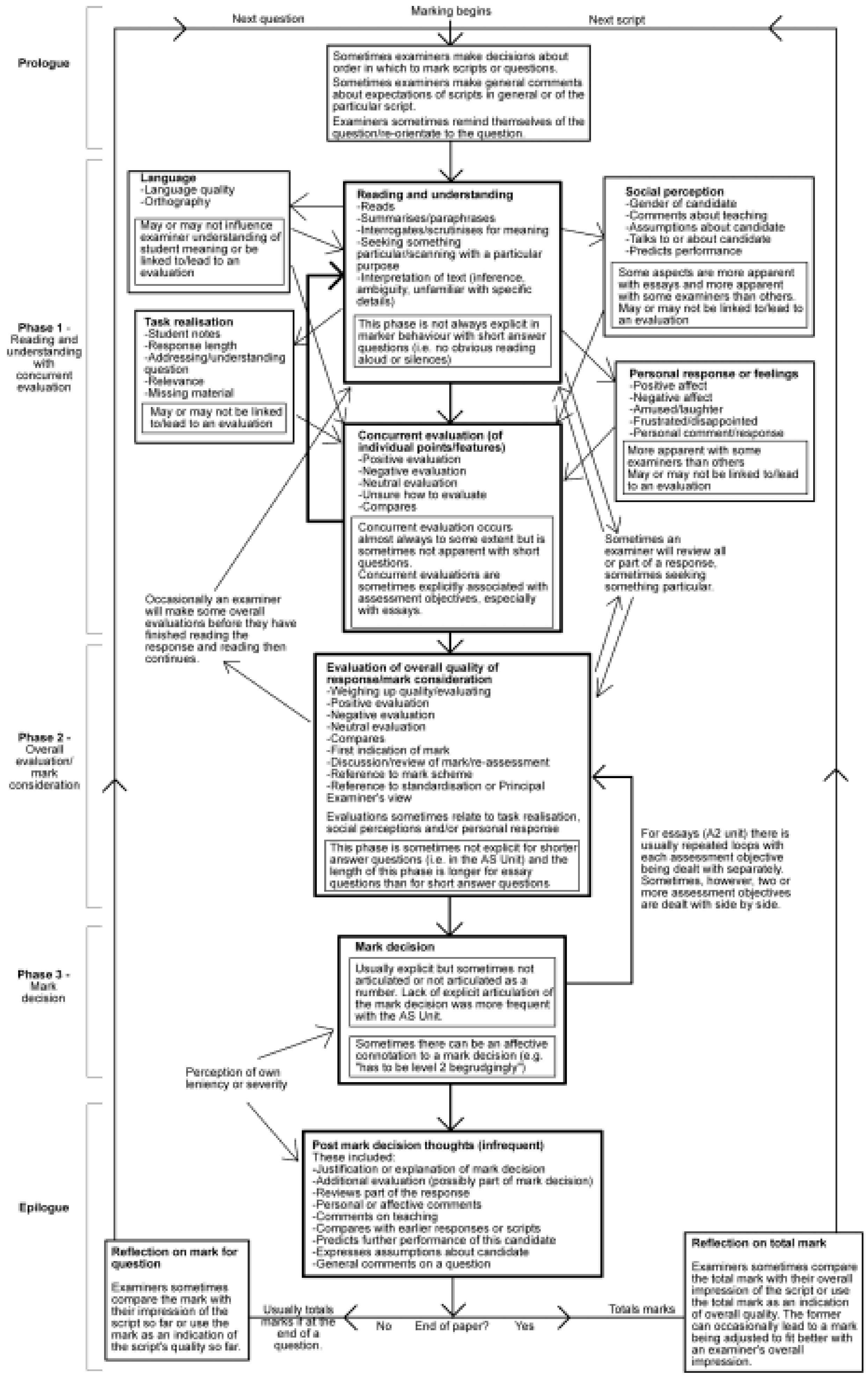





**Figure 1 : (opposite)**
**A tentative model of the marking process in A level geography**

bring together the various aspects of, and influences on, the marking process (text comprehension, cognitive processes, social perceptions, and personal responses) and is compatible with other research in the literature. Variations between marking short-answer questions and marking essays were apparent in certain phases of the model. The phases are outlined briefly below.

### Prologue

When marking begins examiners sometimes make decisions about the order in which to mark scripts or questions and sometimes comment on their expectations of the scripts (e.g. *'surely we will have a good one soon'*) or re-orientate themselves to the question they were about to mark. The prologue occurs fairly infrequently.

**Phase 1 – Reading and understanding with concurrent evaluation and comments on social perceptions of candidates, personal/affective responses and task realisation**

This phase often involves loops of reading and/or paraphrasing parts of the response and then evaluating that part of the response. The process of making sense of the response and making concurrent evaluations tends to be less obvious with short answer questions. Concurrent evaluations are sometimes associated with assessment objectives, especially when marking essays. Reading a student's response can also trigger thoughts regarding the language used by the candidate, task realisation and social and personal responses, and these were sometimes directly associated with, or followed by, a concurrent evaluation.

**Phase 2 – Overall evaluation/mark consideration**

In phase 2 the examiner evaluates the response in a more overall way, possibly weighing up its quality, commenting on strengths and weaknesses. Explicit attempts are likely to be made at this stage to quantify the quality of the response with respect to the mark scheme. The examiner may have initial thoughts as to an appropriate mark and they may consider the advice in the mark scheme and given by the Principal Examiner during standardisation. The evaluations that are made at this stage may relate back to earlier thoughts regarding the task realisation, social perceptions and personal responses that impacted on concurrent evaluations. For the A2 exam, overall evaluations are usually made with regard to each assessment objective in turn and looping occurs between phases 2 and 3.

**Phase 3 – Mark decision**

This phase involves the examiner's decision about the mark. This was usually explicit in protocols but not always, particularly with short answer questions, perhaps because the mark decision occurs quickly and is consequently not articulated. Examiners sometimes reflected on the leniency or severity of their marking when deciding on a mark.

**Epilogue**

Fairly infrequently, additional consideration of the response occurs after the mark decision has been made. This can include, for example, justifying or explaining a mark decision, further evaluation, reviewing part of the response, personal or affective comments, comparisons with other scripts or responses, prediction of further performance by the candidate, and checking whether a total mark matched their overall impression of the script.

The tentative model is illustrated as a flow chart in Figure 1. The model requires further thought and development as well as validation in other subjects and assessments. The interview data were consistent with the coding frame and the proposed model of the marking process.

## Discussion

The findings suggest a number of tentative implications of the research. First, along with the research of Sanderson and others, the current findings emphasise the importance of the process of reading and constructing a full meaning of the student's response as a part of the marking process. The codes 'reads' and 'summarises/paraphrases' were among the most frequently applicable codes and the frequency of reading behaviours seemed to be associated with marker agreement. As well as leading to the unsurprising conclusion that careful reading of responses is important to accurate marking, there may be implications for current moves towards on-screen marking as reading texts on-screen may be more difficult than reading from paper, particularly for longer texts (O'Hara and Sellen, 1997).

Secondly, evaluation processes were very important in the marking process as would be expected. Positive and negative evaluations were among the most commonly observed behaviours. Interestingly, despite the current culture of positive marking, there were fairly similar frequencies of positive and negative evaluations and frequent overlaps of positive and negative evaluations. This is in line with Greatorex's (2000) finding that although mark schemes are designed to positively reward performance with descriptions of performance written in positive terms, examiners' tacit knowledge, perhaps inevitably, leads them to view achievement in both positive and negative ways. Further, lower frequencies of positive evaluations appeared to be associated with severity and with lower marker agreement emphasising the importance of not overlooking positive elements of responses.

Thirdly, comparing a response with other responses seems to be advantageous to marker agreement. Comparisons are to be expected according to Laming (2004) who considers all judgements to be relative. Tversky and Kahneman (1974) suggest that subjects anchor subsequent judgements to initial ones. Indeed, Spear (1997) found that good work was assessed more favourably following weaker material and that high quality work was evaluated more severely following high quality work. Although assessment in UK national examinations usually aspires towards criterion-referenced standards (Baird, Cresswell & Newton, 2000) with the intention that student work is judged against criteria rather than measured by how it compares to the work of others, the findings support the view that it is necessary to have experience with a range of student work in order to understand the criteria fully and to make judgements fairly. Indeed, the importance of using exemplars in the definition and maintenance of standards is generally acknowledged (Wolf, 1995; Sadler, 1987).

The findings of this research support the view that assessment involves processes of actively constructing meaning from texts as well as involving cognitive processes. The idea of examining as a practice that occurs within a social framework is supported by the evidence of some 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 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.

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

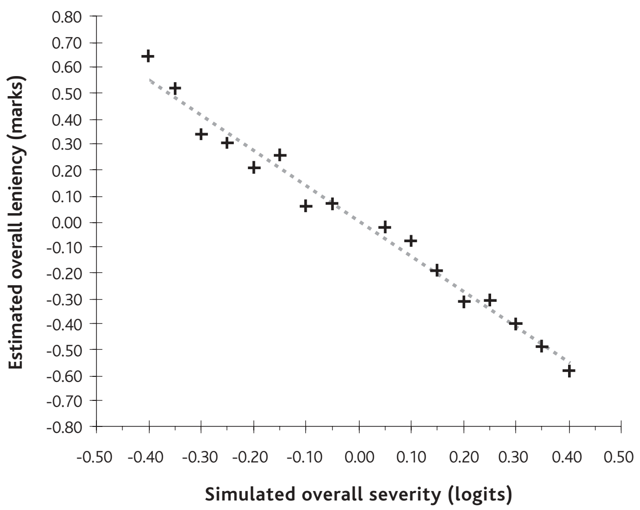
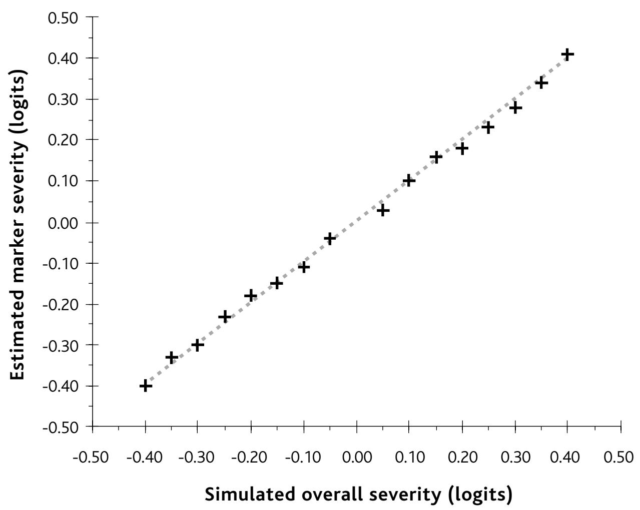



**Figure 2 : Rasch model – estimated severity as a function of simulated severity**



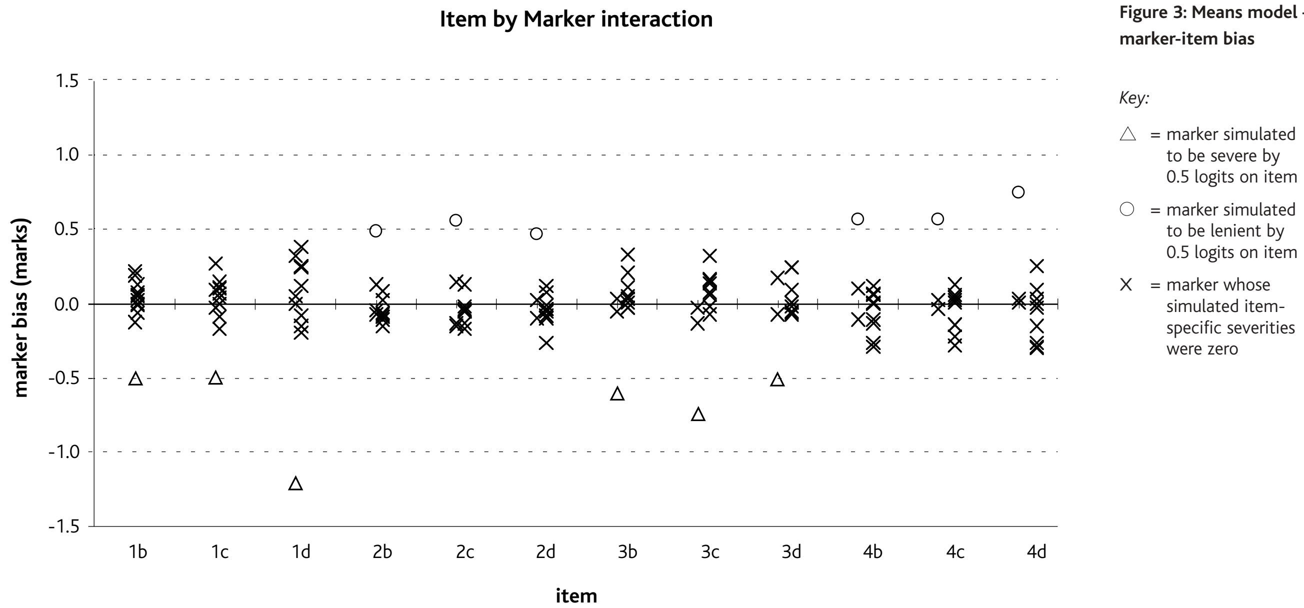




**Figure 3: Means model – marker-item bias**

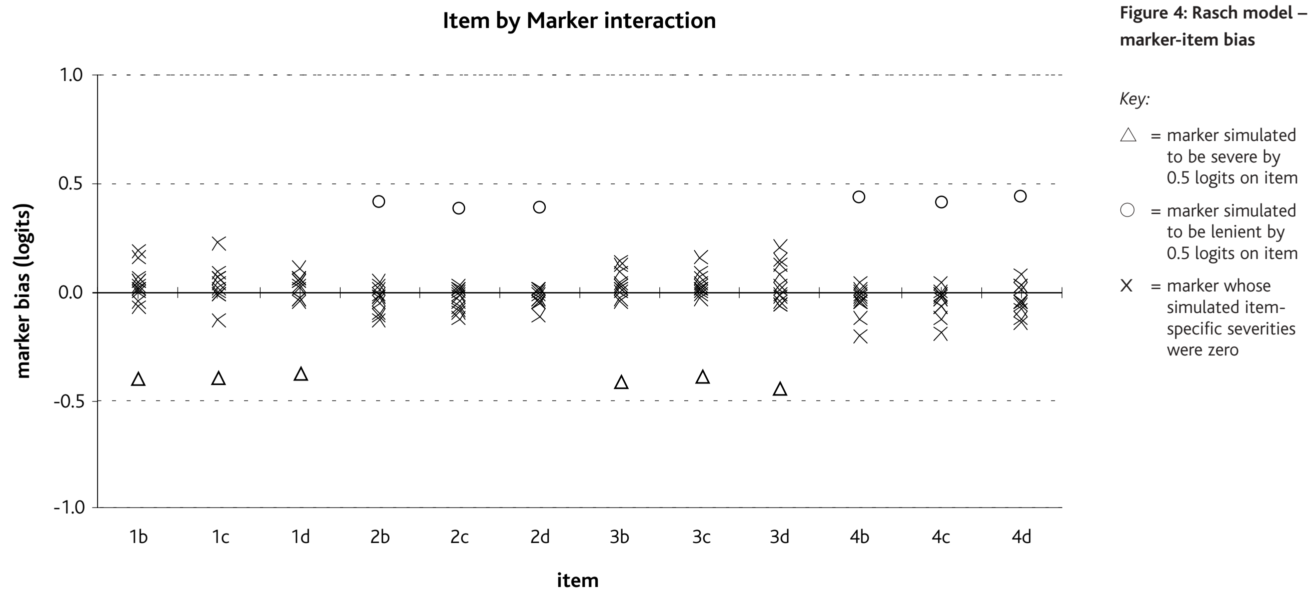



**Figure 4: Rasch model – marker-item bias**

## 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.

# Quantifying marker agreement: terminology, statistics and issues

**Tom Bramley** Research Division

## Introduction

One of the most difficult areas for an exam board to deal with when communicating with the public is in explaining the extent of 'error' in candidates' results. Newton (2005) has discussed this in detail, describing the dilemma facing the exam boards: increased transparency about accuracy of results may lead to decreased public trust in those results and the agencies producing them. Measurement error is often conceptualised as the variability of an individual's score across a set of hypothetical replications (for a critique of the underlying philosophy of this approach, see Borsboom, 2005). In everyday language, this could be presented from the point of view of the candidate as a series of questions:

- Would I have got a different result if I had done the test on a different day?
- Would I have got a different result if the test had contained a different sample of questions?
- Would I have got a different result if the test had been marked by a different person?

I would suggest that whilst all these sources of variability (error) are inherent, it is the third one (marker variability) which is of most concern to the public and the exam board, because it seems to be the one most related to the fairness of the outcome. A great deal of effort goes into standardising all procedural aspects of the marking process and investing in marker training.

The advent of new technologies in mainstream live examinations processing, such as the on-screen marking of scanned images of candidates' scripts, creates the potential for far more statistical information about marker agreement to be collected routinely. One challenge facing assessment agencies is in choosing the appropriate statistical indicators of marker agreement for communicating to different audiences. This task is not made easier by the wide variety of terminology in use, and differences in how the same terms are sometimes used.

The purpose of this article is to provide a brief overview of:

- the different terminology used to describe indicators of marker agreement;
- some of the different statistics which are used;
- the issues involved in choosing an appropriate indicator and its associated statistic.

It is hoped that this will clarify some ambiguities which are often encountered and contribute to a more consistent approach in reporting research in this area.

There is a wide range of words which are often seen in the context of marker agreement, for example: reliability, accuracy, agreement, association, consistency, consensus, concordance, correlation. Sometimes these words are used with a specific meaning, but at other times they seem to be used interchangeably, often creating confusion. In this article I will try to be specific and consistent about usage of terminology. It will already be clear that I have chosen to use 'agreement' as the general term for this discussion, rather than the more commonly used 'reliability'. This is because reliability has a specific technical definition which does not always lead to the same interpretation as its everyday connotation (see section 3).

As might be expected, there are several aspects to marker agreement, and sometimes confusion is caused by expecting a single term (and its associated statistic) to capture all the information we might be interested in. We should be aware that different indicators might be appropriate in different situations. Some considerations which could affect our choice of indicator are listed below:

- Level of measurement – are we dealing with nominal, ordinal or interval-level data?
- Are the data discrete or continuous? (The numerical data is nearly always discrete, but sometimes it is thought to represent an underlying continuum).
- Is there a known 'correct' mark with which we are comparing a given mark or set of marks?
- Are we comparing two markers, or more than two markers?
- How long is the mark scale on the items being compared?
- Where does this marking situation fall on the continuum from completely objective (e.g. multiple-choice item) to subjective (e.g. holistic high-tariff essay)?
- Is the comparison at the level of sub-question, whole question, section, or test?
- What is the intended audience for communicating the information about marker agreement?
- What is the range of situations across which we would like to make generalisations and comparisons?

Rather than attempt an exhaustive survey of all possible combinations of the above factors, I will concentrate on a selection of scenarios which might seem to be most relevant in the context of on-screen marking.

## 1. Objective mark scheme, comparison at sub-question level, low[1] mark tariff (1-3 marks), known correct mark, comparing a single marker

This is probably the most commonly occurring situation. If the mark scheme is completely objective then the correct mark could be determined (in principle) by a computer algorithm. However, I would like to include in this scenario cases where the mark of the Principal Examiner (PE) could legitimately be taken as the 'correct' mark (for example, in applying expert judgment to interpret a fairly objective[2] mark scheme). This scenario should therefore cover the situation which arises in on-screen marking applications where 'gold standard' scripts (where the correct marks on each item are known) are 'seeded' into a marker's allocation of scripts to be marked. I have arbitrarily set the mark limit for this scenario at questions or sub-questions worth up to three marks – a survey of question papers might lead to a better-informed choice.

I would suggest that the best term for describing marker agreement in this scenario is **accuracy**. This is because the correct mark is known. In this scenario, markers who fall short of the desired level of accuracy should be described as 'inaccurate'.

The most informative (but not the most succinct) way to present information collected in this scenario is in an $n \times n$ table like Table 1 below where the rows represent the correct mark and the columns represent the observed mark. The cells of the table contain frequency counts for an individual marker on a particular sub-question or question. This kind of table is sometimes referred to as a 'confusion matrix'.

**Table 1 : Cross-tabulation of frequencies of observed and 'correct' marks on a 3-mark item**

| | | *Observed mark* | | | | |
|---|---|---|---|---|---|---|
| | | *0* | *1* | *2* | *3* | *Row sum* |
| Correct mark | 0 | $n_{00}$ | $n_{01}$ | $n_{02}$ | $n_{03}$ | $n_{0.}$ |
| | 1 | $n_{10}$ | $n_{11}$ | $n_{12}$ | $n_{13}$ | $n_{1.}$ |
| | 2 | $n_{20}$ | $n_{21}$ | $n_{22}$ | $n_{23}$ | $n_{2.}$ |
| | 3 | $n_{30}$ | $n_{31}$ | $n_{32}$ | $n_{33}$ | $n_{3.}$ |
| | Column sum | $n_{.0}$ | $n_{.1}$ | $n_{.2}$ | $n_{.3}$ | N |

The shaded cells are those containing the frequencies of exact agreement between the observed and the correct mark.

The simplest indicator of accuracy would be the overall proportion (or percentage) of raw agreement ($P_0$), which is the proportion of the total frequency coming from the shaded cells.

$$P_0 = \frac{\sum_{0}^{m} n_{ii}}{N}$$

where $m$ is the maximum mark on the question (in Table 1 $m$ = 3).

However, it is likely that we might want to present more information from the data in the cross-table than can be obtained from the single statistic of overall agreement.

For example, we might be interested in whether the observed mark tended to be higher or lower than the correct mark (which might indicate a specific misunderstanding of the mark scheme), and in how far away from the correct mark the observed mark tended to be.

1 The research literature describes many extra statistical possibilities for measuring agreement with dichotomous (1-mark) items, but in the context of examination marking I do not believe there is much to be gained from treating them as anything other than instances of low-tariff items.

2 In practice there can be considerable difficulties in implementing a computer algorithm for marking 'fairly' objective questions – see, for example Sukkarieh *et al*. (2003).

This could be shown by presenting a frequency table of the differences between observed and correct mark. This essentially reduces the $n \times n$ cells in the table of frequencies to a single row of frequencies in the $(2n-1)$ diagonals of the table, as shown in Tables 2 and 3 below.

**Table 2 : Hypothetical data from responses to 90 three-mark gold standard items**

| | | *Observed mark* | | | | |
|---|---|---|---|---|---|---|
| | | *0* | *1* | *2* | *3* | *Row sum* |
| Correct mark | 0 | 11 | 2 | 1 | 0 | 14 |
| | 1 | 4 | 18 | 1 | 0 | 23 |
| | 2 | 1 | 4 | 26 | 2 | 33 |
| | 3 | 1 | 1 | 3 | 15 | 20 |
| | Column sum | 17 | 25 | 31 | 17 | 90 |

Accuracy (overall exact agreement $P_0$) = (11+18+26+15) / 90 = 70 / 90 = 0.778.

**Table 3 : Frequencies of differences between observed and correct mark**

| N | Difference | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|---|
| 90 | Frequency | 1 | 2 | 11 | 70 | 5 | 1 | 0 |

A table in the form of Table 3 would allow the reader to see at a glance:

- how accurate the marker was (relative proportion of cases with zero difference)
- whether the marker tended to be severe (entries with negative numbers) …
- … or lenient (entries with positive numbers)
- the size and frequency of the larger discrepancies.

For completeness, it would be helpful to add a column indicating the total mark for the item, and for comparisons it might be more helpful to show percentages rather than frequencies in the table, as in Table 4 below.

**Table 4 : Percentages of differences between observed and correct mark**

| *N* | *Item max* | *Difference* | *-3* | *-2* | *-1* | *0* | *1* | *2* | *3* |
|---|---|---|---|---|---|---|---|---|---|
| 90 | 3 | % | 1.1 | 2.2 | 12.2 | 77.8 | 5.6 | 1.1 | 0 |

In this form, the table shows the percentage of overall agreement in the highlighted box. For questions worth larger numbers of marks, we might decide to include the boxes either side (or two either side) of the zero difference box in calculating the indicator of accuracy (see section 2).

Is it desirable to summarise the information still further? For routine reporting the above format might still take up too much space. The obvious way to reduce the table further would be simply to summarise the distribution of differences, for example by the mean and standard deviation (SD). It may be that in practice it is difficult to get a feel for the meaning of the SD in this context, and if so the mean absolute difference from the mean (MAD) could be used instead.

**Table 5 : Summary of distribution of differences between observed and correct marks**

| *N* | *Item max* | $P_0$ | *Mean* | *SD* | *MAD* |
|---|---|---|---|---|---|
| 90 | 3 | 0.78 | -0.12 | 0.63 | 0.36 |

Obviously, the more the data are reduced, the more information is lost. Ideally as much information as possible should be preserved in order to facilitate comparisons (for example, between items worth different numbers of marks, between markers who have marked different numbers of items, etc.).

### Other possible statistics

*Kappa*

A more complex statistic than $P_0$ which is often used in this situation is Cohen's Kappa (Cohen, 1960) or weighted Kappa (Cohen, 1968). This indicates the extent of agreement over and above what would be expected by chance. The problem with using it in our context is that we are not really interested in showing that our markers are better than chance at agreeing with the PE, but in finding out how far they fall short of perfection!

A second problem with Kappa is that it is influenced by both the shape of the marginal distributions (i.e. the distribution of row and column totals in the confusion matrix) and the degree to which the raters agree in their marginal distributions (Zwick, 1988). This could be controlled to some extent in a 'gold-standard seeding' scenario by ensuring that equal numbers of pupil responses worth 0,1, 2 and 3 marks were used as the seeds.

However, the verdict of Uebersax (2002a) is that Kappa is only appropriate for testing whether there is more agreement than might be expected by chance, and not appropriate for quantifying actual levels of agreement. A statistic which has attracted so much controversy in the research literature is probably best avoided if the aim is for clear communication.

*Krippendorff's Alpha*

A still more complex statistic is Krippendorff's Alpha (Krippendorff, 2002). This has been designed to generalise to most conceivable rating situations – handling multiple raters, different levels of measurement scale, incomplete data and variable sample size. The same problems apply as for Kappa, with the added disadvantage that the necessary computations are much more long-winded, and do not seem yet to be implemented in standard statistical packages (unlike Kappa). In my opinion it is unlikely that this single statistic could live up to the claims made for it.

*Correlations*

The familiar Pearson product-moment correlation would obviously be inappropriate because it requires continuous data on an interval scale. However, the Spearman rank-order correlation coefficient is also inappropriate (as an indicator of accuracy) because it measures covariation rather than agreement and could thus give misleadingly high values even when exact agreement ($P_0$) was relatively low. This might happen, for example, if the observed mark was consistently one mark higher than the correct mark.

### Summary for scenario 1

The indicator of agreement should be called 'accuracy'.

*N* and $P_0$ should be reported as a minimum, followed by (in order of increasing amount of information):

- mean and SD of differences;
- frequency (%) distribution of differences between observed and correct mark;
- full $n \times n$ cross-table of frequencies.

## 2. Holistic or levels-based mark scheme, high tariff question (10+ marks), single marker compared with team leader or PE

This is another commonly occurring scenario, for example, where a team of markers has been trained to mark a particular essay question. It may be that the PE's mark has a privileged status (i.e. would be given more weight than that of a team member), but it is not necessarily true that the PE's marks are correct. This scenario could also apply where the comparison mark was taken to be the median or mode of several markers, instead of using the PE's mark.

There are several important differences with scenario 1 which need to be taken into account.

First of all, there is (often) assumed to be an underlying continuous trait of quality (or level of performance, or ability), and the responses are assumed to have a 'true' location on this trait. Each marker has their own conceptualisation of this trait, and each response is perceived to lie at a certain position on the trait, this position being a function of the true value, marker-specific effects and residual random error. There is no specifiable algorithm for converting a response (e.g. an essay) into an absolutely correct numerical mark. (This is not the same as saying that there is no rationale for awarding higher or lower marks – the whole point of a well-designed mark scheme and marker training is to provide such a rationale, and to ensure that as far as possible the markers share the same conceptualisation of the trait).

Secondly, although the trait is assumed to be continuous, the marker usually has to award a mark on a scale with a finite number of divisions from zero to the item's maximum. In this scenario with a long mark scale it is often assumed that the marks can be treated as interval-level data.

Thirdly, as mentioned above, it is also often assumed that some kind of random error (again continuous and often assumed to be normally distributed) is an inextricable component of any individual mark.

This means that (even more than with scenario 1) a single statistic cannot capture all the relevant information about marker agreement. This is because markers can differ in:

1. their interpretation of the 'true' trait (i.e. what is better and what is worse);
2. severity / leniency (a systematic bias in the perceived location of the responses on the trait);
3. scale use (a different perception of the distribution of the responses on the trait);
4 'erraticism' – the extent to which their marks contain random error.[3]

3 Conceptually, erraticism can be distinguished from differences in interpretation of the 'true' trait by considering the latter as differences *between* markers in where they perceive the response to lie on the trait, whereas erraticism is differences *within* a marker as to where they perceive the same response to lie on hypothetical replications of marking. In practice, these two are difficult to separate.



There is less likely to be a 'correct' mark in this scenario, and gold standard items are less likely to be used because of the time investment in creating and using them. However, there may well be occasions where a single marker's mark on a set of items needs to be compared with those of a senior marker (I will assume a PE for brevity), whose marks can be treated as the correct marks.

In this case it is possible to use the same approach as scenario 1, but just to concentrate on the distribution of differences between the marker and the PE. With a 15-mark item, the differences would need to be grouped into ranges – seven 'bins' seems a reasonable number[4], as shown in Table 6 below (which uses the same percentages as Table 4).

**Table 6 : Example distribution of differences between a marker and the PE on a 15-mark item**

| *N* | *Item max* | *Difference* | *≤ -8* | *-7 to -5* | *-4 to -2* | *-1 to +1* | *+2 to +4* | *+5 to +7* | *≥ +8* |
|---|---|---|---|---|---|---|---|---|---|
| 90 | 15 | % | 1.1 | 2.2 | 12.2 | 77.8 | 5.6 | 1.1 | 0 |

Again, the percentage of cases in the bin containing zero (the highlighted box) could form one indicator of agreement. It might be less appropriate to refer to this as accuracy – perhaps simply **agreement** is a better term for this kind of agreement. My suggestion for terminology for the agreement statistic in this case would be '$P_{agr1}$' which would be interpreted as 'the proportion of cases with agreement between marker and PE within a range of ±1 mark'.

As in scenario 1, this distribution could further be reduced to the mean and SD of differences.

**Table 7 : Mean and SD of mark differences between marker and PE**

| *N* | *Item max* | *Mean* | *SD* |
|---|---|---|---|
| 90 | 15 | -0.8 | 2.39 |

If we were prepared to assume that the differences were normally distributed (this could be checked graphically) then we could infer from the data in Table 7 that the marker was on average 0.8 (≈1) mark below the PE and that approximately 95% of the time their mark was between 6 marks below and 4 marks above that of the PE (these are the rounded mark points ± 2 SDs either side of the mean of –0.8). If we did not want to make assumptions about the shape of the distribution of differences it might be preferable to report the mode or median and the interquartile range (IQR), instead of the mean and SD.

The mean (or median) difference indicates the severity or lenience in the marker's marks, and the SD (or IQR) of the differences indicates the extent to which the marker's interpretation of the trait disagrees with the PE's and/or their degree of erraticism. Is there a way to extend this approach to assess differences in scale use between markers?

One solution is to plot the difference between marker and PE against the PE's mark, as shown in Figure 1. Any patterns in this plot would reveal differences in scale usage – for example, Figure 1 shows that this marker was on average about 1 mark severe, but less so at the low end of the mark range and more so at the high end. These differences could be highlighted by fitting a smoothed line to the points.

4 If the mark scheme is levels-based, there may be a natural 'bin' corresponding to the range of marks awarded in each level.

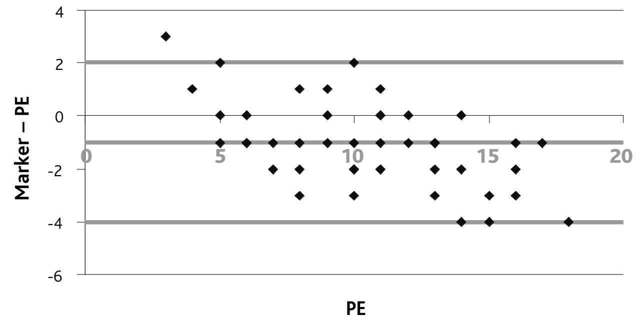



**Figure 1 : Difference between marker and PE's mark (on a hypothetical 20-mark essay) plotted against PE's mark**

The broad lines superimposed on the plot show the mean difference, and two SDs above and below the mean. Altman and Bland (1983, 1986) recommend this kind of graph as providing the most informative visual display of differences between two measurement instruments or methods purporting to measure the same quantity[5].

It might be argued that if a plot is to be produced it would probably be easier for non-experts to interpret a simple plot of marker's mark against PE's mark (Figure 2). If the plot contained an identity line (i.e. the line representing where the marker and PE's marks would be identical) then inspection of this plot could reveal all the types of differences discussed above:

- if the points do not lie in a straight line this indicates that the marker and PE perceive the trait differently – the lower the correlation the greater this difference;
- if the points tend to lie above (or below) the identity line this indicates lenience (or severity);
- if the points tend to be above the identity line at low marks and below at high marks, or vice versa, (or if any other non-linear patterns are observed) this indicates different scale use.

Note that the dotted line in Figure 2 is not a best-fit regression line, but the identity line.

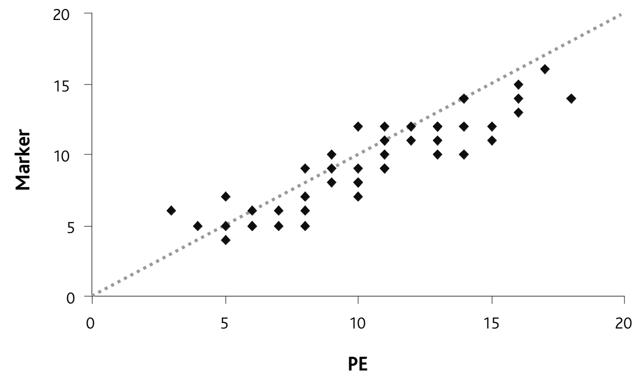



**Figure 2 : Marker's mark plotted against PE's mark (on the hypothetical 20-mark essay)**

5 Their paper was in a medical measurement context, where the question of interest was 'can instrument 2 be used interchangeably with instrument 1?' I would argue that this is a reasonable analogy for this scenario 2 context where we want to know the extent to which the marker is interchangeable with the PE. (They used the average of the two instrument readings for the x-axis of their graph, but I have used the PE's mark for simplicity).

Although it might be easier for non-experts to comprehend data presented in the form of Figure 2, Altman and Bland (1983) argue that plots like Figure 1 are preferable, for the following reasons:

- much of the space in a plot like Figure 2 will be empty (as this example illustrates well);
- the greater the range of marks, the greater the agreement will appear to be;
- the less highly correlated the data, the more difficult it becomes to perceive severity/lenience and different scale use by visual inspection.

A similar approach could also be used in a situation where there is multiple marking of a set of responses. Each marker could be compared against the average mark (or average of the other markers excluding their own mark) instead of against the PE. However, such situations are unlikely to arise outside a research exercise because of the costs involved in multiple marking.

Comparisons of the marker agreement statistics from this scenario with those from other situations are possible, but should be made with caution. In particular, it is important to allow for any differences in the length of the mark scale. It may also be necessary to specifically select samples of responses which cover the full mark range in order to detect any differences in scale use[6]. Comparisons will only be valid if the situations compared have used similar schemes for sampling responses.

### Other possible statistics

*Correlation*

The correlation coefficient is a very widely used statistic, often seen in this context. It indicates the extent of linear association between two variables and thus could legitimately be used to show the extent to which the marker and PE share the same concept of what is better and what is worse. (This has been referred to as 'consistency' by some authors, e.g. Stemler, 2004). However, it cannot tell us anything about relative severity/lenience or scale use. Also, it requires there to be some variability in both sets of marks. Although in an ideal situation we would seek to ensure an appropriate range of marks, the 'mean difference' method described above does not require this. We could still produce the distribution of differences between marker and PE if all the responses had received the same mark from the PE – but the correlation between marker and PE in such a case would be zero. It should also be noted that a high value for the correlation coefficient can mask some fairly large differences – for example, the correlation in the data displayed in Figure 2 is 0.90, but Figure 1 shows that there are several cases where the absolute difference between marker and PE was three marks or more.

*Regression*

It is possible to summarise the data in graphs like Figure 2 by a regression equation of marker's mark (y) on PE's mark (x). This is essentially fitting the model:

$y = a + bx + e$

where $a$ is the intercept, $b$ is the slope, and $e$ is random error.

The extent to which the regression line differs from the identity line could be assessed by testing whether $a$ is significantly different from 0 and $b$ is significantly different from 1.

6 This will also help to mitigate any floor and ceiling effects when interpreting differences between marker and PE.

This regression approach has yet to convince me of its worth. The slope parameter $b$ confounds the correlation and the SD ratio of the two sets of marks, and both parameters might be more sensitive to sample size and outliers in the data than the simple mean of the differences would be. Also, for the results to apply more generally the responses should be sampled at random. Altman and Bland (1983) only recommend the use of regression in the context of prediction, not comparison. However, other researchers may feel that this approach has more to recommend it than I have suggested.

### Summary for scenario 2

The indicator of agreement should be called 'agreement'.

The PE's mark has been used as the comparison mark in scenario 2 for brevity, but this could be replaced by the average of a group of markers in a multiple-marking scenario.

If a single indicator is to be used, $P_{agrN}$ has been suggested, which is the proportion of scripts with a difference between marker and PE in a ±$N$-mark range around zero. $N$ could be increased as the total mark for the question (or sub-test or test) increases.

For fuller diagnosis of the different kinds of differences between marker and PE, the distribution of differences between their marks should be examined:

- The higher the SD, the more they perceived the trait differently, or the more their marks contained random error.
- The more positive (or negative) the mean, the more lenient (or severe) the marker compared to the PE.
- Scatter plots of the difference between marker's mark and PE's mark versus PE's mark can reveal differences in perceived distribution of responses on the trait, in addition to the above two points.

## 3. Reliability of marking

The previous scenarios have concentrated on methods for assessing a single marker's performance in terms of agreement with the correct mark on an objective item (scenario 1), and agreement with the PE's mark on a more subjective item (scenario 2). The term 'reliability' has been deliberately avoided. I would suggest we do not talk about the reliability of an individual marker, but reserve the term 'reliability' for talking about a set of marks. Thus reliability is a term which is perhaps best applied to an aggregate level of marks such as a set of component total scores.

The definition of reliability comes from what has been called 'true score theory', or 'classical test theory' (see, for example, Lord and Novick, 1968). The key point to note is that reliability is defined as the ratio of true-score variance to observed score variance. This very specific technical definition means that it is easy for non-experts to be misled when they read reports about reliability. Reliability refers to a set of scores, not to an individual score. The size of the variance ratio (which can range from 0 to 1) depends on the true variability in the sample. If there is no true score variance, all the observed differences will be due to error and the reliability coefficient will be zero – so the size of the reliability coefficient depends both on the test and on the sample of pupils taking the test.

### Cronbach's Alpha

There are several ways of estimating test reliability, which vary depending on what the source of the errors is deemed to be. One commonly used index of reliability is Cronbach's Alpha (Cronbach, 1951). One way of viewing this statistic is that it treats the individual item responses (marks) as repeated 'ratings' of the same pupil. The proportion of the total variance due to inter-item covariance estimates the reliability[7]. Alpha is referred to as 'internal consistency reliability' because it indicates the extent to which the items are measuring the same construct – or in other words the extent to which pupils who are above (or below) the mean on one item are above (or below) the mean on other items.

Applying the same reasoning to the situation where we have pupils with papers marked by the same set of markers, we can see that Cronbach's Alpha could be applicable here. The total scores from the different markers are the repeated ratings. The reliability of marking would be the proportion of total variance due to differences between pupils. Alpha would indicate the extent to which pupils who were above the mean according to one marker were above the mean according to the other markers – what we might term 'inter-marker consistency reliability'.

However, it is important to note that the size of this statistic would not be affected by systematic differences in severity or leniency between the markers. Adding or subtracting a constant number of marks from every value for a single marker would not change the size of Cronbach's Alpha. This type of marker consistency reliability could only be obtained from a situation where multiple markers had marked the same set of responses, and thus is likely to be more useful in research exercises than in 'live' monitoring of markers.

### Intraclass correlations and general linear models

Cronbach's Alpha can be viewed as a special case of what are known as 'intraclass correlations' or ICCs (Shrout and Fleiss, 1979). These statistics are all based on analysis of variance, and are thus (in my opinion) difficult to communicate to non-specialists. Choosing the appropriate version of the ICC for the given data is of critical importance and should be done by a statistician. It is possible to choose a version of the ICC which is sensitive to differences in both consistency (correlation) and absolute agreement (von Eye and Mun, 2005). Some see this as an advantage, others as a disadvantage (Uebersax, 2003). Most versions of the ICC require double or multiple marking of the same set of responses.

Intraclass correlations themselves arise in more general linear modelling techniques such as generalizability theory (e.g. Cronbach *et al.*, 1972) and multilevel modelling (e.g. Snijders and Bosker, 1999). Approximate global indices of reliability can be derived from these more complex analyses. In fact, one of the main motivations for the development of generalizability theory was to enable the magnitude of different sources of variability in the observed score (e.g. that due to different markers) to be estimated.

### Standard error of measurement

Once a reliability coefficient has been estimated it is possible to derive a standard error of measurement, SEM (see, for example, Harvill, 1991). An approximate 95% confidence interval for the observed score around a given true score is given by ±2 SEMs. These standard errors are arguably easier to interpret than reliability coefficients (which are ratios of variances) because they can be treated as distances in units of marks and thus can be compared to other meaningful mark ranges such as a grade band, or the effective range of observed scores. They are less sample dependent than the reliability coefficient, and can also be generated from generalizability theory and from Rasch (and IRT) modelling.

7 The formula for Cronbach's Alpha also contains an adjustment factor of N/(N-1) to allow it to range between 0 and 1.

### Multi-facet Rasch models

An alternative to the general linear model would be to fit a multi-facet Rasch model (Linacre, 1994). This approach is described by Stemler (2004) as providing a 'measurement estimate' of marker agreement, because the severities/leniencies of the markers are estimated jointly with the abilities of the pupils and difficulties of the items within a single frame of reference – reported as an equal-interval logit scale. Analysis of marker fit statistics can identify 'misfitting' markers who perceived the trait differently from the other markers. Myford and Wolfe (2003, 2004) show that it is possible to use the output from a many-facet Rasch analysis to diagnose other rater effects such as central tendency (overusing the middle categories of the mark scale), a 'halo' effect (a tendency to award similar marks to the same candidate on different questions) and differential severity/leniency (a tendency to be severe or lenient towards particular subsets of candidates).

Both these approaches (general linear models and multi-facet Rasch models) are statistically complex, generating many statistical indicators which can test different hypotheses about individual markers or groups of markers. The indicators from different analyses (i.e. on different sets of data) are unlikely to be comparable. However, both approaches can be used (with certain assumptions) in situations where the script is split into item response groups which are allocated to different markers, without the need for multiple marking of the same responses, which means that both methods are feasible options in some on-screen marking environments.

### Summary for scenario 3

- The term 'reliability' should be reserved for use in its technical sense as a ratio of variances.
- Intraclass correlations are appropriate for reporting reliability, but different ICCs are applicable in different data collection scenarios, and expert statistical advice is essential.
- Where possible, it is preferable to report standard errors of measurement rather than reliability coefficients.
- General linear models and multi-facet Rasch models can diagnose many different aspects of rater agreement. Statistics generated from one set of data are unlikely to be directly comparable with those generated from another.

## Conclusion

The choice of a statistical indicator of marker agreement depends on the situation and reporting purpose. I have argued that simple statistics, based on the distribution of differences between marker and correct mark, or marker and PE, are the easiest to interpret and communicate.

> *A study that reports only simple agreement rates can be very useful; a study that omits them but reports complex statistics may fail to inform.* (Uebersax, 2002b)

It will be interesting to see whether exam boards pick up the gauntlet thrown down by Newton (2005) and risk the short-term cost in terms of public trust by becoming readier to report indices of marker agreement. If they do, it will be important to choose indices which reveal more than they conceal. This last point is well illustrated by Vidal Rodeiro (2007, this issue) – the reader is encouraged to compare in her article tables 4 and 11 with tables 5, 6 and 12.

**References**


Altman, D.G. & Bland, J.M. (1983). Measurement in medicine: the analysis of method comparison studies. *The Statistician*, **32**, 307–317.

Bland, J.M. & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, **i**, 307–310.

Borsboom, D. (2005). *Measuring the mind: conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

Cohen, J. (1968). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213–220.

Cronbach, L.J. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika*, **16**, 297–334.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: theory of generalizability of scores and profiles*. New York: Wiley & Sons.

Eye, A. von & Mun, E.Y. (2005). *Analyzing rater agreement: manifest variable methods*. Mawhah, NJ.: Lawrence Erlbaum Associates.

Harvill, L.M. (1991). An NCME Instructional Module on Standard Error of Measurement. *Educational Measurement: Issues and Practice*, **10**, 2, 33–41.

Krippendorff (2002). Computing Krippendorff's Alpha-Reliability. http://www.asc.upenn.edu/usr/krippendorff/webreliability2.pdf. Accessed January 2006.

Linacre, J.M. (1994). *Many-Facet Rasch Measurement*. 2nd Edition. Chicago: MESA Press

Lord, F.M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Myford, C.M. & Wolfe, E.W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part 1. *Journal of Applied Measurement*, **4**, 4, 386–422.

Myford, C.M. & Wolfe, E.W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part 2. *Journal of Applied Measurement*, **5**, 2, 189–227.

Newton, P.E. (2005). The public understanding of measurement inaccuracy. *British Educational Research Journal*, **31**, 4, 419–442.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, **86**, 2, 420–428.

Snijders, T.A.B., & Bosker, R.J. (1999). *Multilevel analysis. an introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Stemler, S.E. (2004). A comparison of consensus, consistency and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, **9**, 4. Retrieved from http://PAREonline.net/getvn.asp?v=9&n=4 November, 2005.

Sukkarieh, J.Z., Pulman, S. G. & Raikes, N. (2003). *Auto-marking: using computational linguistics to score short, free text responses*. Paper presented at the 29th conference of the International Association for Educational Assessment (IAEA), Manchester, UK.

Uebersax, J. (2002a). Kappa coefficients. http://ourworld.compuserve.com/homepages/jsuebersax/kappa.htm Accessed 22/01/07.

Uebersax, J. (2002b). Raw agreement indices. http://ourworld.compuserve.com/homepages/jsuebersax/raw.htm Accessed 22/01/07.

Uebersax, J. (2003). Intraclass Correlation and Related Methods http://ourworld.compuserve.com/homepages/jsuebersax/icc.htm Accessed 22/01/07.

Vidal Rodeiro, C. L. (2007). Agreement between outcomes from different double marking models. *Research Matters: A Cambridge Assessment Publication*, **4**, 28–34.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, **103**, 3, 374–378.


ASSURING QUALITY IN ASSESSMENT

# Agreement between outcomes from different double marking models


**Carmen L. Vidal Rodeiro** Research Division


## Introduction

The practice of arranging for students' work to be marked by more than one person is a subject of great interest in educational research (see, for example, Cannings *et al*. 2005, Brooks, 2004, White, 2001 or Partington, 1994). However, deciding if double marking is worthwhile incorporates a perennial dilemma. Intuitively, it seems to increase the reliability of the assessment and shows fairness in marking, but this needs to be proven a benefit in order to justify the additional time and effort that it takes. Awarding bodies struggle to recruit enough examiners to mark scripts once, never mind twice, and therefore double marking of all examination papers can be a difficult task.

In the context of GCSE or GCE examinations, double marking can be a means to enhance the reliability of the marking process. One of the principal concerns of any examination board is to ensure that its examinations are marked reliably. It is essential that each examiner is applying the same standard from one script to the next and that each examiner is marking to the same standard as every other examiner. Although Pilliner (1969) had demonstrated that reliability increases as the size of the marking team increases, it was Lucas (1971) who observed that the greatest improvement came from increasing the size of the marking team from one to two and that additional benefits derived from using teams of three or more markers were of smaller magnitude.



## Double marking models

Double marking is more common in examinations where the assessment is known to be subjective, for example, examinations involving writing an essay. In these cases, the main methods of double marking are:

a. **Blind double marking.** The first marker makes no annotations on the work being marked and the second marker examines all pieces of work as left by students.

b. **Non-blind or annotated double marking.** In this case, the first marker makes annotations on the work being marked and the second marker marks it with this information known. This may involve varying degrees of information available to the second marker, for example, annotations to draw attention to points in the text or marks written on answers.

Whatever method is used for double marking examinations, there must be a method of resolving differences between markers. Some of the methods that can be employed for this task are:

a. Discuss and negotiate the marks on all the differences or on specified differences.

b. Take the mean of the marks. This may be done for all differences or for specified differences. However, there are studies that suggest that taking the average of two marks is not the best way to reconcile the differences. For example, Massey and Foulkes (1994) suggested that the average of two blind marks may not always be a sound estimate. It remains at least arguable that the greater the difference between two markers the more likely it is that one has seen something the other has not.

c. Resort to a third marker, who could mark the script afresh or, based on the two previous marks, produce a final mark.

## Aim of the research

The main purpose of this research is to evaluate the agreement between marks from different double marking models, in particular, blind and annotated double marking. We focus on agreement concerning total marks across questions in the examination paper (or component) concerned. We acknowledge that future technologies may change the current marking practice so that instead of one examiner marking the whole of a candidate's paper, questions might be allocated individually to different examiners.

Specific aims are:

1. To measure marking outcomes and agreement between first and second marking.
2. To compare second marking workload in relation to the double marking models, including the impact of examiner experience.
3. To measure reconciliation workload (number required plus time taken).

## Data and methods

### Description of the data and the task

Two General Certificate of Secondary Education (GCSE) units, OCR English and OCR Classical Greek, were selected for this study. For English, one component was chosen: Literary Heritage and Imaginative Writing, Higher Tier. The total number of marks for this unit was 40. For Classical Greek, the component 2, Verse Literature, was selected. The total number of marks for this unit was 60. For each subject, a two hundred script sample from the June 2004 examination was retained.

Five examiners per subject were invited to participate in this research: a principal examiner (PE), two senior examiners (or experienced assistant examiners) and two assistant examiners.

For both English and Classical Greek, the scripts were split into two packs of one hundred scripts. Each assistant examiner was allocated one hundred scripts from a range of different marks. These scripts had all marks and marking annotations removed. Each of the more experienced or senior examiners was allocated two packs of scripts. One pack contained one hundred scripts that had the marks and comments from the original examiners on them, whereas for the one hundred scripts in the other pack, these were removed. In each pack the scripts were from a range of different marks. We ensured that each script appeared in only one pack.

For each subject, the examiners were asked to mark the scripts following the same marking instructions that had been used in the original marking of the examination. A meeting with the examiners took place before the re-marking started. In the meeting, the principal examiners reviewed the mark scheme with the assistant and senior examiners in order to identify any marking issues. It should be noted that this meeting was not a full standardisation meeting and that, as this research was done under experimental conditions, some of the quality assurance procedures that are carried out during live marking were not performed.

Reconciliation was carried out when the difference between the original 'live' mark and the mark awarded in this study for the same script exceeded 10% of the mark range. The principal examiners in each subject performed this task, producing a final mark.

After the marking and the reconciliation were performed, the experiment produced four marking outcomes in each subject:

1. Original: 200 scripts with the original marks awarded in the June 2004 session.

*Plus re-marking of the same 200 scripts using three different strategies:*

2. Treatment 1: Blind re-marking by two assistant examiners (marking 100 scripts each) plus the reconciliation by the PE as needed.
3. Treatment 2: Blind re-marking by two senior (or experienced) examiners (marking 100 scripts each) plus the reconciliation by the PE as needed.
4. Treatment 3: Non-blind or annotated re-marking by two senior (or experienced) examiners (marking 100 scripts each) plus the reconciliation by the PE as needed.

### Statistical methodology

There is little consensus about what statistical methods are best to analyse markers' agreement. There are many alternatives in the literature although the most commonly used are the correlation coefficients and the Kappa statistics (see Uebersax, 2003, for an overview of the different statistics that are used in this field and Bramley, 2007, for a discussion of how they might be applied in a double marking context).

*Correlation coefficients*

Usually, the first step in this type of analysis is to plot the data and draw the line of equality on which all points would lie if the two markers gave exactly the same mark every time. The second step is to calculate the correlation coefficient between the two markers ($\rho$) which measures the degree to which two variables are linearly related. When the relationship between the two variables is nonlinear or when outliers are present, the correlation coefficient incorrectly estimates the strength of the relationship. Plotting the data before computing a correlation coefficient enables the verification of a linear relationship and the identification of potential outliers.

On the principle of allowing for some disagreement but not too much, in the context of double marking examinations Wood and Quinn (1976) proposed that between-marker correlations in the region of 0.50 to 0.60 would seem to be realistic.

*Measures of agreement*

Early approaches to the study of markers' agreement focussed on the observed proportion of agreement, that is, the proportion of cases in which the markers agreed. However, this statistic does not allow for the fact that a certain amount of agreement can be expected on the basis of chance alone. A chance-corrected measure of agreement, introduced by Cohen (1960), has come to be known as Kappa. For two markers, it is calculated as follows:

$$\kappa = \frac{P_a - P_c}{1 - P_c},$$

where $P_a$ is the proportion of marks in which the markers agree and $P_c$ is the proportion of marks for which agreement is expected by chance.

Table 1 shows the degree of agreement for different values of Kappa (Landis and Koch, 1977). The limits of this classification are arbitrary and can vary according to the study carried out. Kappa can take negative values if the markers agree at less than chance level and it can be zero if there is no agreement greater or lesser than chance.

**Table 1 : Degree of agreement and values of Kappa**

| *Degree of agreement* | *Kappa* |
|---|---|
| Excellent | ≥ 0.81 |
| Good | 0.80 – 0.61 |
| Moderate | 0.60 – 0.41 |
| Poor | 0.40 – 0.21 |
| Bad | 0.20 – 0.00 |
| Very bad | < 0.00 |

## Results

Examiners were able to mark, on average, 5 or 6 scripts per hour. This did not seem to vary whether the scripts were annotated or blind. Some examiners originally thought that marking the annotated ones would be swifter but this proved not to be the case. There seems to be no difference between the time employed by assistant and senior examiners in marking their scripts.

### GCSE Classical Greek

Table 2 displays summary statistics of the marks awarded in the different marking treatments. The means do not differ very much and the standard deviations are very similar in all cases. The marks given to the scripts are all rather high (the minimum available mark for the component is 0 and the lowest mark awarded by an examiner is 17). The re-markers appear very similar in their overall marks but all mark, on average, more generously than the original markers.

**Table 2 : Summary statistics of the marks awarded in the different marking treatments**

| | *N* | *Mean* | *Standard Deviation* | *Minimum* | *Maximum* |
|---|---|---|---|---|---|
| Original | 200 | 43.72 | 9.06 | 17 | 60 |
| Treatment 1 | 200 | 44.05 | 8.82 | 17 | 59 |
| Treatment 2 | 200 | 43.93 | 9.15 | 15 | 60 |
| Treatment 3 | 200 | 44.09 | 9.19 | 18 | 60 |

Table 3 displays the absolute (unsigned) differences between the original marks and the three sets of re-marks. The average mark change between the original and the first treatment (blind re-mark by assistant examiners) is bigger than for the other treatments. The smallest value corresponds to the non-blind re-mark (treatment 3). This last difference is probably caused by seeing the actual marks awarded but it might, in part, be due to comments providing additional insight into why the original examiner awarded a particular mark.

**Table 3 : Absolute differences in marks**

| | *Mean Difference* | *Standard Deviation* |
|---|---|---|
| Original – Treatment 1 | 2.16 | 1.69 |
| Original – Treatment 2 | 1.97 | 1.73 |
| Original – Treatment 3 | 0.67 | 0.84 |

The simplest way to describe agreement would be to show the proportion of times two markers of the same scripts agree, or the proportion of times two markers agree on specific categories. Table 4 displays these proportions.

The percentages of exact agreement between the original marks and the different sets of re-marks are 16%, 17% and 50%. When agreement is widened to include adjacent marks, agreement levels increase. For example, for treatment 1 (blind re-marking by assistant examiners) the marks differ by no more than +/- one in around 43% of the scripts marked and by +/- three marks in around 78% of the scripts. For treatment 2 (blind re-marking by senior or more experienced assistant examiners) the marks differ by +/- one in around 45% of the scripts marked and by +/- three in around 87% of the scripts. For treatment 3 (non-blind re-marking) the marks differ by +/- one mark in around 87% of the scripts marked and in three or fewer marks in around 98% of the scripts.

**Table 4 : Distribution of differences between original and experimental marks**

| *Difference in marks* | *Treatment 1 (%)* | *Treatment 2 (%)* | *Treatment 3 (%)* | *Total (%)* |
|---|---|---|---|---|
| ≤ -6 | 3.5 | 1.0 | 0.0 | 1.5 |
| -5 | 2.0 | 2.5 | 0.0 | 1.5 |
| -4 | 7.5 | 4.0 | 1.5 | 4.3 |
| -3 | 10.5 | 9.0 | 2.0 | 7.2 |
| -2 | 10.0 | 16.0 | 7.0 | 11.0 |
| -1 | 11.5 | 14.5 | 26.5 | 17.5 |
| 0 | 16.0 | 17.0 | 50.0 | 27.7 |
| 1 | 15.0 | 13.0 | 11.0 | 13.0 |
| 2 | 10.5 | 9.5 | 2.0 | 7.3 |
| 3 | 4.5 | 8.0 | 0.0 | 4.2 |
| 4 | 6.0 | 2.5 | 0.0 | 2.8 |
| 5 | 1.0 | 1.5 | 0.0 | 0.8 |
| ≥ 6 | 2.0 | 1.5 | 0.0 | 1.2 |

Table 4 provides, again, evidence that removing previous marks and comments from scripts does make a difference. There are alternative interpretations of this. A negative perspective would suggest that examiners who are asked to re-mark scripts cannot help but be influenced by the previous judgements, however much they try to ignore them and form their own opinion. A positive view would argue that the non-blind re-markers can see why the original mark was awarded and are happy to concur; even though had they marked blind they might well not have spotted features noted by the original examiner.

Pearson's correlation coefficients between marking treatments are displayed in Table 5.

**Table 5 : Pearson's correlation coefficients**

| | *Original* | *Treatment 1* | *Treatment 2* | *Treatment 3* |
|---|---|---|---|---|
| Original | 1.0000 | 0.9538 | 0.9588 | 0.9940 |
| Treatment 1 | 0.9538 | 1.0000 | 0.9478 | 0.9554 |
| Treatment 2 | 0.9588 | 0.9478 | 1.0000 | 0.9639 |
| Treatment 3 | 0.9940 | 0.9554 | 0.9639 | 1.0000 |

The correlation coefficients are high (the smallest correlation appears between the original mark and treatment 1: $\rho = 0.9538$) and of an order which would normally be regarded as an indicator of high reliability of marking. The highest correlation appears between the original marks and the non-blind re-marks. The correlation between the treatment 2 (blind re-mark by senior or more experienced assistant examiners) and the original marks is higher than the correlation between treatment 1 and the original marks, which might reflect the relative experience of the examiners.

Another way of assessing the agreement between pairs of markers is the use of Kappa (Kappa statistics are displayed in Table 6). Again, this table provides confirmation of the hypothesis that the marking of two examiners would be affected by whether or not previous marks and comments had been removed from the scripts.

**Table 6 : Kappa statistics**[1]

| | *Original* |
|---|---|
| Treatment 1 | 0.7609 |
| Treatment 2 | 0.8103 |
| Treatment 3 | 0.9327 |

1 The values of the Kappa statistic provided in this table were not obtained using the formula given in this article but using an extended version (Cohen, 1968).

*Reconciliation*

Using the 10% criterion described in the methodology section, we determined which scripts needed reconciliation. For Classical Greek the maximum and minimum marks are 60 and 0, respectively. Then, if for a particular script, the absolute difference between two marks is bigger than 6, the script needs reconciliation and this is undertaken by the principal examiner. Table 7 displays the numbers and percentage (in brackets) of scripts that needed reconciliation.

**Table 7 : Scripts that needed reconciliation**

| | | *Examiner Experience* | | *Marking* | |
|---|---|---|---|---|---|
| | *Total* | *Blind Assistant* | *Blind Senior/ Experienced* | *Blind* | *Non-blind* |
| Reconciliation | 16 (2.7) | 11 (5.5) | 5 (2.5) | 16 (4.0) | 0 (0.0) |

Only 16 of the re-marked scripts needed reconciliation (2.7%). Of those, 11 scripts were blind re-marked by assistant examiners and 5 by senior or more experienced assistant examiners. This confirms that when experienced examiners re-marked scripts the differences with the original marks are smaller than when assistant examiners did so. Non-blind re-marked scripts did not need reconciliation.

Only 4 of the reconciliation outcomes correspond with the mean of two prior markings, although 12 of the reconciliation outcomes are within +/- two marks of this mean. Reconciliation for blind marks by assistant examiners produces outcomes more widely distributed around the mean of prior marking than for senior or more experienced assistant examiners (see Table 8). Note that the numbers involved in the reconciliation task are too small to draw any strong conclusions.

**Table 8 : Difference between the mean of two marks and the reconciliation outcome**

| | *Treatment 1* | | *Treatment 2* | |
|---|---|---|---|---|
| *Difference* | *Frequency* | *%* | *Frequency* | *%* |
| -3 | 1 | 9.1 | 0 | 0.0 |
| -2 | 0 | 0.0 | 1 | 20.0 |
| -1 | 1 | 9.1 | 2 | 40.0 |
| 0 | 2 | 18.2 | 2 | 40.0 |
| 1 | 1 | 9.1 | 0 | 0.0 |
| 2 | 3 | 27.3 | 0 | 0.0 |
| 3 | 2 | 18.2 | 0 | 0.0 |
| 4 | 1 | 9.1 | 0 | 0.0 |

Reconciling differences is likely to prove better than averaging because it takes better advantage of the information available or even gathers and uses some more. However, this approach might be difficult to transfer to large scale public examinations. The fact that non-blind re-marking required no reconciliation may well be an important advantage in large scale operations.

During the reconciliation task, the principal examiner 'reconciled' around five scripts per hour. If we had changed the cut-off point for reconciliation and reconciled scripts where the absolute difference between two marks was bigger than 3 (5% of the mark range) then the time employed and the cost that it entailed would have made the reconciliation task much more expensive. The total percentage of scripts needing reconciliation would have been around 12%. 17.5% of the blind re-marked scripts and 1.5% of the non-blind re-marked scripts would have had to be reconciled.

### GCSE English scripts

Table 9 displays summary statistics of the marks awarded in the different marking treatments. The mean is half a mark lower in treatment 1 (blind re-mark by assistant examiners) and three marks higher in treatment 2 (blind re-mark by senior examiners). With regard to treatment 3 (non-blind re-mark), the mean is quite close to the original, being only half a mark higher. The standard deviation of the re-marks is smaller than the one in the original marks. The minimum and the maximum marks are similar in all marking treatments.

**Table 9 : Summary statistics of the marks awarded in the different marking treatments**

| | *N* | *Mean* | *Standard Deviation* | *Minimum* | *Maximum* |
|---|---|---|---|---|---|
| Original | 200 | 22.08 | 7.78 | 4 | 40 |
| Treatment 1 | 200 | 21.53 | 6.89 | 5 | 38 |
| Treatment 2 | 200 | 25.31 | 7.31 | 6 | 40 |
| Treatment 3 | 200 | 22.73 | 7.62 | 4 | 39 |

Table 10 displays the absolute differences between the original marks and the three treatments. The average mark change between the original scripts and treatment 1 is 4.49. For treatment 2, the mean is 5.64, which is bigger than for the other treatments. The smallest value corresponds to the non-blind marking (third treatment), where the minimum difference, 0, was achieved in 71 cases. This table provides confirmation of the hypothesis that the marking of two examiners would be affected by whether or not previous marks and comments had been removed from the scripts. Annotations might affect to what exactly within an answer a subsequent examiner will pay attention. Something marked up by the first examiner might be emphasised to a second examiner when they might not have noticed it themselves and, if the first examiner missed something salient, the second examiner may be more likely to do so too (Wilmut, 1984).

**Table 10 : Absolute differences in marks**

| | *Mean Difference* | *Standard Deviation* |
|---|---|---|
| Original – Treatment 1 | 4.49 | 3.68 |
| Original – Treatment 2 | 5.64 | 4.19 |
| Original – Treatment 3 | 1.84 | 2.20 |

The percentages of exact agreement between the original marks and the different sets of re-marks are 8%, 3% and 36%, respectively. Figures in Table 11 provide evidence of much wider disagreement (in total marks) between English examinations than between Classical Greek examinations. This is no doubt related to the nature of the English examination questions, which are much less tightly structured, allowing for greater freedom in composing a response and requiring more subjective judgement by markers.

**Table 11 : Distribution of differences between original and experimental marks**

| *Difference in marks* | *Treatment 1 (%)* | *Treatment 2 (%)* | *Treatment 3 (%)* | *Total (%)* |
|---|---|---|---|---|
| < -13 | 1.0 | 4.5 | 0.0 | 1.8 |
| -13 to -11 | 1.5 | 5.5 | 0.0 | 2.3 |
| -10 to -8 | 6.5 | 15.0 | 2.5 | 8.0 |
| -7 to -5 | 8.0 | 15.0 | 5.0 | 9.3 |
| -4 to -2 | 18.5 | 23.0 | 20.5 | 20.7 |
| -1 | 7.5 | 5.5 | 15.5 | 9.5 |
| 0 | 8.0 | 2.5 | 35.5 | 15.3 |
| 1 | 8.5 | 6.5 | 7.5 | 7.5 |
| 2 to 4 | 19.5 | 11.1 | 9.5 | 13.4 |
| 5 to 7 | 10.0 | 6.5 | 3.0 | 6.5 |
| 8 to 10 | 6.0 | 4.0 | 1.0 | 3.7 |
| 11 to 13 | 3.0 | 1.0 | 0.0 | 1.3 |
| > 13 | 2.0 | 0.0 | 0.0 | 0.7 |

Figure 1 illustrates the marks awarded in the three different treatments and the original marks and Pearson's correlation coefficients are displayed in Table 12.

Figure 1 permits a comparison to be made between the marks awarded to the scripts in the different treatments. It can be seen that the variations between the markers' judgements were considerably reduced when they were marking scripts with the original marks and comments on them.

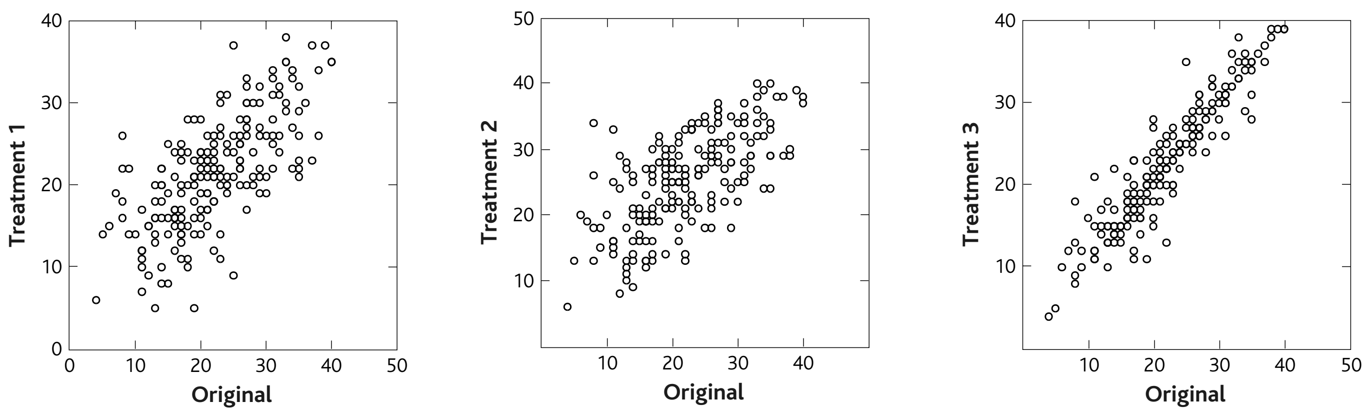



**Figure 1 : Scatter diagrams illustrating the relationship between the marks awarded in the different treatments**



The correlation coefficients with the original marks are not very high for treatments 1 and 2 indicating that, to a certain extent, the re-markers do not agree closely with the original marks. They also do not agree with one another. The highest correlation appears between the original marks and treatment 3. The correlation between treatment 2 and treatments 1 and 3 is higher than the correlation between treatment 2 and the original marks.

**Table 12 : Pearson's correlation coefficients**

| | *Original* | *Treatment 1* | *Treatment 2* | *Treatment 3* |
|---|---|---|---|---|
| Original | 1.0000 | 0.6951 | 0.6593 | 0.9346 |
| Treatment 1 | 0.6951 | 1.0000 | 0.6789 | 0.7417 |
| Treatment 2 | 0.6593 | 0.6789 | 1.0000 | 0.7276 |
| Treatment 3 | 0.9346 | 0.7417 | 0.7276 | 1.0000 |

In terms of the Kappa statistic, for the first treatment we obtain a moderate agreement with the original marks (0.4908). For the second treatment, the value of Kappa, 0.4371, indicates moderate to poor agreement. The level of agreement is higher for treatment 3, with a value of Kappa of 0.7783 (similar to the blind re-mark in Classical Greek), which is a sign of a good agreement.

*Reconciliation*

Scripts needing reconciliation were determined using the 10% criterion. In this case, reconciliation is performed if the difference in marks is bigger than 4. Table 13 displays the numbers and percentage (in brackets) of scripts that needed reconciliation.

**Table 13 : Scripts that needed reconciliation**

| | *Total* | *Examiner Experience* | | *Marking* | |
|---|---|---|---|---|---|
| | | *Blind Assistant* | *Blind Senior/ Experienced* | *Blind* | *Non-blind* |
| Reconciliation | 202 (33.7) | 76 (38.0) | 103 (51.5) | 179 (44.8) | 23 (11.5) |

In English, the number of scripts needing reconciliation was much higher than for Classical Greek. 202 of the re-marked scripts needed reconciliation. Among them, 76 scripts were blind re-marked by assistant examiners and 103 by senior examiners. 23 scripts that were non-blind re-marked needed reconciliation.

In the three treatments, reconciliation generally provides different outcomes than averaging two marks (see Table 14) and increases the correlation with the original marks and the blind re-marking. Cresswell (1983) demonstrated that the simple addition of the two markers' scores will rarely produce a composite score with the highest reliability possible.

If we had reduced the cut-off point for reconciliation to +/- 2 marks (5% of the mark range) then the reconciliation task would have become enormous. The total percentage of scripts needing reconciliation would have been around 50%. 64% of the blind re-marked scripts and 22% of the non-blind re-marked scripts would have had to be reconciled, greatly increasing costs.

**Table 14 : Difference between the mean of two marks and the reconciliation outcome**

| | *Treatment 1* | | *Treatment 2* | | *Treatment 3* | |
|---|---|---|---|---|---|---|
| *Difference* | *Frequency* | *%* | *Frequency* | *%* | *Frequency* | *%* |
| -6 | 1 | 1.3 | 2 | 1.9 | 1 | 4.3 |
| -5 | 1 | 1.3 | 5 | 4.8 | 3 | 13.0 |
| -4 | 2 | 2.6 | 4 | 3.9 | 3 | 13.0 |
| -3 | 4 | 5.3 | 12 | 11.6 | 1 | 4.3 |
| -2 | 13 | 17.1 | 16 | 15.5 | 0 | 0.0 |
| -1 | 10 | 13.2 | 10 | 9.7 | 0 | 0.0 |
| 0 | 11 | 14.8 | 12 | 11.6 | 0 | 0.0 |
| 1 | 12 | 15.8 | 5 | 4.8 | 4 | 17.4 |
| 2 | 8 | 10.5 | 14 | 13.6 | 5 | 21.7 |
| 3 | 10 | 13.2 | 15 | 14.6 | 4 | 17.4 |
| 4 | 1 | 1.3 | 3 | 2.9 | 1 | 4.3 |
| 5 | 2 | 2.6 | 4 | 3.9 | 1 | 4.3 |
| 6 | 0 | 0.0 | 1 | 1.0 | 0 | 0.0 |
| 7 | 1 | 1.3 | 0 | 0.0 | 0 | 0.0 |

## Conclusions and discussion

A first conclusion that can be drawn from this study is that there is a contrast between Classical Greek and English, the former being more reliably marked. Newton (1996) found the same type of contrast between Mathematics, traditionally the most reliably marked subject, and English.

Although in Classical Greek some of the questions required relatively extended answers, the task of the examiners was to award a mark for a specified response. In English, the examiners' task was generally to evaluate the quality of the work. This involved more interpretation and therefore more scope for a difference in opinion.

The results of this investigation appear to provide evidence that removing previous marks and comments from scripts does make a difference. It would seem that examiners who are asked to re-mark scripts cannot help but be affected by previous judgements: the non-blind re-markers can see why the original mark was awarded and they might be happy to concur. Had the second examiners marked blind, they might well not have spotted features noted by the original examiners but also might have spotted features not noted by the original examiners. However, had they marked non-blind, they might have been influenced by incorrect marks or annotations.

There is a need for further research into non-blind double marking. It is necessary to be sure that the second marker will always have the confidence to reject the influence of the marking or the annotations.

One serious impediment to double marking is the increase in administrative time and costs which it entails. Feasibility is a further issue due to the shortage of specialist markers in the UK.

Finally, it should be pointed out that the marking carried out in this research is done under experimental conditions. In the live marking of the examinations, a standardisation meeting is held in order to establish a common standard that is used to maintain the quality of marking during the marking period. Although in this research a meeting with the examiners took place before the re-marking and the principal examiners reviewed the mark schemes with the examiners in order to identify any marking issues, there was no full standardisation meeting. Also, in the live marking period, when the examiners are submitting their marks there are a series of quality control procedures, for example, monitoring the marking, adjustments to the marks of an individual examiner or clerical checks (details on the quality assurance procedures can be found in QCA Code of Practice, 2005). In this research we examined the marks without performing these procedures.


**References**

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication*, **4**, 22–28.

Brooks, V. (2004). Double marking revisited. *British Journal of Educational Studies*, **52**, 29-46.

Cannings R., Hawthorne K., Hood K. & Houston H. (2005). Putting double marking to the test: a framework to assess if it is worth the trouble. *Medical Education*, **39**, 299-308.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 7-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213- 220.

Cresswell M.J. (1983). *Optimum weighting for double marking procedures*. AEB Research Report, RAC281.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159-174.

Lucas, A.M. (1971). Multiple marking of a matriculation Biology essay question. *British Journal of Educational Psychology*, **41**, 78-84.

Massey, A. & Foulkes, J. (1994). Audit of the 1993 KS3 Science national test pilot and the concept of quasi-reconciliation. *Evaluation and Research in Education*, **8**, 119-132.

Newton, P. (1996). The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal*, **22**, 405-420.

Partington, J. (1994). Double marking students' work. *Assessment and Evaluation in Higher Education*, **19**, 57-60.

Pilliner, A.E.G. (1969). Multiple marking: Wiseman or Cox? *British Journal of Educational Psychology*, **39**, 313-315.

QCA (2005). *GCSE, GCE, VCE, GNVQ and AEA Code of Practice 2005/2006*. London: Qualifications and Curriculum Authority.

Uebersax, J. (2003). *Statistical Methods for the Analysis of Agreement*. http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm. Accessed April 2006.

White, R. (2001). Double marking versus monitoring examinations. *Philosophical and Religious Studies*, **1**, 52-60.

Wilmut, J. (1984). *A pilot study of the effects of complete or partial removal of marks and comments from scripts before re-marking them*. AEB Research Report RAC315.

Wood, R. & Quinn, B. (1976). Double impression marking of English language essays and summary questions. *Educational Review*, **28**, 229-246.


ASSURING QUALITY IN ASSESSMENT

# Item-level examiner agreement


**Nicholas Raikes and Alf Massey** Research Division



## Abstract

Studies of inter-examiner reliability in GCSE and A-level examinations have been reported in the literature, but typically these focused on paper totals, rather than item marks. See, for example, Newton (1996). Advances in technology, however, mean that increasingly candidates' scripts are being split by item for marking, and the item-level marks are routinely collected. In these circumstances there is increased interest in investigating the extent to which different examiners agree at item level, and the extent to which this varies according to the nature of the item.

Here we report and comment on intraclass correlations between examiners marking sample items taken from GCE A-level and IGCSE examinations in a range of subjects.

The article is based on a paper presented at the 2006 Annual Conference of the British Educational Research Association (Massey and Raikes, 2006).


## Introduction

One important contribution to the reliability of examination marks is the extent to which different examiners' marks agree when the examiners mark the same material. Without high levels of inter-examiner agreement, validity is compromised, since the same mark from different examiners cannot be assumed to mean the same thing. Although high reliability is not a sufficient condition for validity, the reliability of a set of marks limits their validity.

Research studies have in the past investigated inter-examiner reliability, but typically these focussed on agreement at the level of the total mark given to scripts. The operational procedures followed by examination Boards for documenting examiner performance also often involve recording details of discrepancies between examiners at the script total level. New technologies are facilitating new ways of working with examination scripts, however. Paper scripts can now be scanned and the images transmitted via a secure Internet link to examiners working on a computer at home. Such innovations are creating an explosion in the amount of item-level marks available for analysis, and this is fostering an interest in the degree of inter-examiner agreement that should be expected at item level. The present article provides data that will help inform discussions of this issue.

## The source of our data

The analysis presented in the present article was of data collected during trials of new ways for examiners to record item-level marks. All marking for the trials was done using paper scripts (i.e. no marking was done on screen, the only innovation was in the way the markers recorded their marks). The marks therefore give an indication of the kind of agreement that can be expected between examiners marking whole scripts on paper. The results are indicative only because the study marking was low stakes for the examiners (i.e. no candidate's result depended on the marks and the examiners knew their performance would not be appraised), and also because different methods of recording marks were being trialled, which might have had a small effect on their reliability.

The five components for which data were available are as follows:

- **IGCSE Foreign Language French: Listening**
  Multiple choice (m/c) and short answer textual answers worth 1 or 2 marks
- **IGCSE Development Studies: Alternative to Coursework**
  Short answers worth 1–6 marks
- **A-level Chemistry: Structured Questions**
  m/c and short answers worth 1–5 marks
- **A-level Economics: Data Response and Case Study**
  Short, textual answers worth 1–6 marks; some longer textual answers worth 8–12 marks
- **A-level Sociology: Principles and Methods**
  Candidates chose 2 from 6, 25-mark essay items

## Inter-examiner agreement at script-total level

Although item-level data are the main focus of our article we present results for script totals in Table 1. 'ITR per mark' in Table 1 is the Implied Time Restriction per mark, equal to the time allowed for the examination divided by the maximum mark available for the examination. The column labelled 'ICC $r_{totals}$' gives the intraclass correlation coefficient between the examiners' total marks for the scripts. The intraclass correlation may be interpreted as the proportion of variance in the set of marks that is due to the candidates (i.e. after examiner effects have been controlled for). That is, if there is perfect agreement between the examiners on every script, the intraclass correlation coefficient will be 1; but if there is no agreement and the marks appear random, the coefficient will be 0. Bramley (2007) discusses approaches to quantifying agreement between pairs of examiners in this issue of *Research Matters*, but correlation based measures are useful when considering the relationship between more than two examiners, as is the case here.

**Table 1 : Intraclass correlations for script totals**

| *Subject* | *Max mark* | *Time (mins)* | *ITR per mark* | $N_{examiners}$ | $N_{scripts}$ | *ICC* $r_{totals}$ |
|---|---|---|---|---|---|---|
| French | 48 | 45 | 0.9 | 4 | 300 | 0.995 |
| Dev. Stud. | 35 | 90 | 2.6 | 4 | 265 | 0.917 |
| Chemistry | 60 | 60 | 1.0 | 3* | 298 | 0.992 |
| Economics | 50 | 110 | 2.1 | 4 | 294 | 0.774 |
| Sociology | 50 | 90 | 1.8 | 3* | 252 | 0.863 |

* One Chemistry and one Sociology examiner dropped out of the trials

Looking first at the Implied Time Restrictions per mark in Table 1, it seems that the question paper designers generally gave candidates about 1 minute per mark for papers consisting of multiple choice and short answer questions, and about 2 minutes per mark for papers involving more extended answers. Development Studies was apparently generous in the amount of time given to candidates, since this question paper only contained short answer questions.

All the ICCs in Table 1 are high, indicating a considerable degree of agreement between the examiners. As might be expected, the agreement was highest for the French and Chemistry papers, consisting of multiple choice and short answer questions, a little lower for Development Studies, containing only short answer questions, and a little lower still for Sociology, consisting solely of 25-mark essays. It is slightly surprising that the Economics examiners showed the lowest levels of agreement, given that the Economics paper contained some short answer questions. However, as discussed below, the ICC for Economics does not appear low when the Implied Time Restriction is taken into account.

There is a striking relationship between the Implied Time Restriction per mark and ICC. If Development Studies with its apparently generous time restriction is excluded, the Pearson correlation between these two quantities is -0.99 – that is, the degree of agreement between examiners at script-total level for these four question papers can be almost perfectly predicted from the Implied Time Restriction per mark.

## Inter-examiner agreement at item level

We classified items as 'objective', 'points' or 'levels' according to the kind of marking required as follows:

- **Objective marking** – items that are objectively marked require very brief responses and greatly constrain how candidates must respond. Examples include items requiring candidates to make a selection (e.g. multiple choice items), or to sequence given information, or to match given information according to some given criteria, or to locate or identify a piece of information (e.g. by marking a feature on a given diagram), or to write a single word or give a single numerical answer. The hallmark of objective items is that all credit-worthy responses can be sufficiently pre-determined to form a mark scheme that removes all but the most superficial of judgements from the marker.
- **Points based marking** – these items generally require brief responses ranging in length from a few words to one or two paragraphs, or a diagram or graph, etc. The key feature is that the salient points of all or most credit-worthy responses may be pre-determined to form a largely prescriptive mark scheme, but one that leaves markers to locate the relevant elements and identify all variations that deserve credit. There is generally a one-to-one correspondence between salient points and marks.
- **Levels based marking** – often these items require longer answers, ranging from one or two paragraphs to multi-page essays or other extended responses. The mark scheme describes a number of levels of response, each of which is associated with a band of one or more marks. Examiners apply a principle of best fit when deciding the mark for a response.

Tables 2 to 4 present data about inter-examiner agreement at item level. Looking first at the bottom right hand cell of each table, the overall mean

**Table 2 : means and standard deviations of ICCs for OBJECTIVE items**

| | Mean ICC (Objective items) (N$_{items}$) Standard Deviation of the ICCs | | | | |
|---|---|---|---|---|---|
| *Max mark* | *French* | *Dev. Stud.* | *Chemistry* | *Economics* | *All* |
| **1** | **0.975** (21) *0.027* | **0.981** (1) | **0.950** (3) *0.073* | **0.978** (1) | **0.972** (26) *0.033* |
| **2** | - | **0.978** (1) | - - | - - | **0.978** (1) |
| **6** | **0.986** (1) | - | - | - | **0.986** (1) |
| **All** | **0.975** (22) *0.027* | **0.980** (2) *0.002* | **0.950** (3) *0.073* | **0.978** (1) | **0.973** (28) *0.032* |

**Table 3 : means and standard deviations of ICCs for POINTS items**

| | Mean ICC (Points items) (N$_{items}$) Standard Deviation of the ICCs | | | | |
|---|---|---|---|---|---|
| *Max mark* | *French* | *Dev. Stud.* | *Chemistry* | *Economics* | *All* |
| **1** | **0.877** (15) *0.082* | **0.883** (2) *0.044* | **0.837** (25) *0.090* | - | **0.854** (42) *0.086* |
| **2** | **0.852** (3) *0.058* | **0.609** (4) *0.156* | **0.885** (12) *0.062* | **0.774** (2) *0.149* | **0.817** (21) *0.138* |
| **3** | - | **0.719** (4) *0.082* | **0.899** (3) *0.049* | **0.517** (3) *0.034* | **0.712** (10) *0.165* |
| **6** | - | **0.809** (1) | - | **0.548** (1) | **0.679** (2) *0.185* |
| **All** | **0.873** (18) *0.078* | **0.717** (11) *0.143* | **0.856** (40) *0.082* | **0.608** (6) *0.147* | **0.820** (75) *0.126* |

ICC is, as expected, highest for the objective items (0.973), next highest for the points items (0.820), and lowest for the levels items (0.773).

Table 2 shows the objective items were marked very reliably regardless of the subject or the maximum mark available (though most of the objective items were on the French Listening paper and only two were worth more than one mark).

One-mark points items (top row of Table 3) were marked a little less reliably than one-mark objective items (top row of Table 2), as expected. The right-most column of Table 3 shows that overall, mean ICC for the points items decreased with rising maximum mark. Surprisingly, this trend does not apply within all the subjects. For Chemistry, the only subject with a considerable number of items worth more than one mark, there is a rising trend.

The six 25-mark Sociology essay items (near the bottom right of Table 4) marked using a levels marking scheme were marked very reliably (average ICC = 0.825, with little variation between the items). It is not

**Table 4 : means and standard deviations of ICCs for LEVELS items**

| | Mean ICC (Levels items) (N$_{items}$) Standard Deviation of the ICCs | | | |
|---|---|---|---|---|
| *Max mark* | *Dev. Stud.* | *Chemistry* | *Economics* | *All* |
| **4** | **0.890** (1) | - | - | **0.890** (1) |
| **8** | - (1) | **0.740** | - | **0.740** (1) |
| **10** | - (1) | **0.567** | - | **0.567** (1) |
| **12** | - (1) | **0.585** | - | **0.585** (1) |
| **25** | - | - | **0.825** (6) *0.044* | **0.825** (6) *0.044* |
| **All** | **0.890** (1) | **0.631** (3) *0.095* | **0.825** (6) *0.044* | **0.773** (10) *0.115* |

obvious why there was less inter-examiner agreement for the Economics levels items, though the Economics examiners also had the lowest overall mean ICC for the points items. The Sociology results show it is possible to have lengthy pieces of extended writing marked reliably.

## Conclusion

In this article we have provided some detailed information about inter-examiner agreement levels that were obtained from IGCSE and A-level examiners marking whole scripts on paper in a non-live context from examinations in five subjects.

Intraclass correlation (ICC) coefficients generally indicated a high degree of agreement between examiners at both script total and item level. When items were classified according to their marking schemes as 'objective', 'points' or 'levels', the objective items were on average marked more reliably than the points items, which were on average marked more reliably than the levels items, as expected. On average reliability decreased with rising maximum mark for points items, but surprisingly this trend was reversed for Chemistry. Six 25-mark Sociology essay questions marked against a levels mark scheme were marked very reliably, proving that it is possible to achieve high reliability for essay marking.

We found a very strong relationship between the Implied Time Restriction (ITR) per mark[1] that was imposed on candidates and the intraclass correlation (ICC) obtained for script total marks. A Pearson correlation of -0.99 was found between ITR per mark and ICC when one subject, IGCSE Development Studies, which had an apparently long ITR per mark, was excluded from the calculation. Implied Time Restriction per mark therefore appears to be a useful indicator of the level of inter-examiner agreement that should be expected at total script mark level.

1 The Implied Time Restriction per mark equals the time allowed for an examination divided by the maximum mark available for the examination, i.e. it is the average time available to candidates for earning a mark.


**References**

Bramley, T. (2007). Quantifying marker agreement; terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication*, **4**, 22–28.

Massey, A.J. & Raikes, N. (2006). *Item-Level Examiner Agreement*. Paper presented at the 2006 Annual Conference of the British Educational Research Association, 6–9 September 2006, University of Warwick, UK.

Newton, P.E. (1996). The Reliability of Marking of General Certificate of Secondary Education Scripts: Mathematics and English. *British Educational Research Journal*, **22**, 4, 405–420.


CAMBRIDGE ASSESSMENT NETWORK

# Fostering communities of practice in examining


**Andrew Watts** Cambridge Assessment Network


*This is a shortened version of a paper given at the International Association for Educational Assessment Conference in May 2006.*

## The necessity of communities of practice in a judgemental system

The term 'community of practice', when applied to examining in a traditional system, is usually used to denote the system of induction, cooperative working, supervision and development of examiners that aims to overcome the error to which their judgements are prone. Dylan Wiliam wrote in 1996 that 'maintenance of standards requires that those responsible for setting standards are full participants in a community of practice, and are trusted by the users of assessment results'. His observation does not only apply to assessments of school attainment. Alison Wolf (1995), writing about competence-based assessment, describes how assessors 'operate in terms of an internalised, holistic set of concepts'. With examples from a number of educational and vocational contexts she concludes '… how important and, potentially, how effective assessor networks are. They are, in fact, the key element in ensuring consistency of judgement' (p.77).

## Subjectivity and objectivity

It has been common to characterise the judgements made in assessment as 'subjective' in contrast to more automated assessments which are 'objective'. Pierre Bourdieu (1990) however, in his analyses of social practice, calls any division between these two concepts 'artificial' and particularly argues against the privileging of an 'objective' standpoint. Sueellen Shay (2005) applies Bourdieu's analysis to the case of a university Engineering Department's assessment of undergraduates' final year theses, which she describes as 'complex tasks'. She describes such assessments within the logic of social practice and asserts that 'all judgement is both objectively and subjectively constituted'. She writes that this kind of professional judgement requires 'a double reading … an iterative movement'. From an objective perspective, assessors can 'observe, measure and map reality independent of the representations of those who live in it'. Subjectively, on the other hand, assessment is 'an embodiment of the assessor'; it is 'relational', 'situational', 'pragmatic' and 'sensitive to the consequences of [the] assessment'. Such 'double readings' enable the judges to assess a 'socially constituted, practical mastery' (p.675).

Shay's concept of a socially based 'double reading' presents us with a *requirement* for assessment to take place within a community of practice. Thus, assessment is understood within a social theory of learning, such as Wenger's (1998), which recognises the place of components like 'community, identity, meaning and practice' (p.5). This supports the view that a balancing of subjective and objective perspectives should be sought in making assessments, and that the community of practice provides an appropriate context for the assessment of complex tasks.

## Reliability and the use of new technologies

Concern for greater reliability has motivated the search for more automated ways of managing and marking examination scripts. Paper scripts can be scanned and the images transmitted via a secure Internet connection to markers working on a computer at home. There is then the potential for all examiners to mark the same training scripts online, and for a Team Leader to call up instantly any script that an examiner wishes to discuss with them. Team Leaders may more closely monitor and support examiners during marking, since all marked scripts, together with the marks and annotations, are instantly available. Standardising scripts, with marks already agreed by senior examiners, can be introduced 'blind' into online marking allocations to check that examiners have not drifted from the common standard, and statistical methods for flagging potential aberrant marking may be employed. All these procedures may improve the reliability of marking, but they might also undermine the argument for maintaining a community of practice amongst all examiners. If the bulk of examiners can be trained and effectively monitored online, do they need to come together at all?

## Validity as a prime concern

Shay (2004) describes assessment as a 'socially situated interpretive act'. She argues that validation of the assessment is what matters crucially and that the on-going process of evaluating the soundness of our interpretations is a community process. She quotes Bernstein, stating that validation requires 'the existence of a community of enquirers who are able, willing and committed to engage in the argumentation'. She argues that the 'typical technologies of our assessment and moderation systems … privilege reliability' and we fail to use these technologies as 'opportunities for dialogue about what we really value as assessors, individually and as communities of practice' (p.676).

In a paper delivered to the first Cambridge Assessment Conference in October 2005, Alison Wolf noted that 'very often we discuss assessment as an essentially technical affair'. We pursue reliability and lose sight of broader issues like the limitations of what we are testing and the effect of our assessments on those being assessed.

## Validity and communities of practice

Wenger's (1998) description of the concept of communities of practice is a dissertation on human learning. Its most challenging thoughts concerning assessment do not refer to the way examiners should learn their trade but to the conditions in which true learning might take place. He says that school curricula, in order to make the process of learning orderly and manageable, often 'reify' the process and thus decrease the possibility that learning which is committed and involved might take place. This can then result in only a limited kind of learning being assessed. Wenger concludes:

> *[such learning] can be misleading in that evaluation processes reflecting the structure of a reified curriculum are circular. Students with a literal relation to a subject matter can reproduce reified knowledge without attempting to gain ownership of its meaning. An evaluation process will become more informative regarding learning that has actually taken place to the extent that its structure does not parallel that of instruction too closely, but instead conforms to the structure of engagement in actual practice and the forms of competence inherent in it*. (p. 265)

Whether the performance of a candidate in an assessment 'conforms to the structure of engagement in actual practice' in a domain of knowledge will be something, as we noted in Shay's comments above, that only members of a community of practice will be able to judge. It will therefore be essential that, in the coming changes to assessment practice, the importance of fostering these groups is not overlooked.


**References**

Bourdieu, P. (1990). *The Logic of Practice*. Stanford, CA: Stanford University Press.

Shay, S. (2004). The assessment of complex performance: a socially-situated interpretive act. *Harvard Educational Review*, **74**, 3, 307–329.

Shay, S. (2005). The Assessment of Complex Tasks: a double reading. *Studies in Higher Education*, **30**, 6, 663–679.

Wenger, E. (1998). *Communities of Practice: Learning, Meaning and Identity*. Cambridge, UK: Cambridge University Press.

Wiliam, D. (1996). Standards in Examinations: a matter of trust? *The Curriculum Journal*, **7**, 3, 293.

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

Wolf, A. (2005). *What can we measure? And what should we be measuring?* Paper delivered to the Cambridge Assessment Conference. October 17 2005.






# Research News

## Conferences and seminars

### Professor James Flynn seminar

IQ scores have been going up since they were first recorded but does that mean people are becoming more intelligent? This question was debated by Professor James Flynn at a seminar in December hosted by the Psychometrics Centre at Trinity College, Cambridge. Professor Flynn's presentation was followed by a discussion led by Neil Mackintosh, Fellow of King's College and Professor of Experimental Psychology, Cambridge, and John White, Professor of the Philosophy of Education, University of London.

### UK Rasch Users' Group

In February members of the Assessment Research and Development Division attended a one day conference in Cambridge of the UK Rasch Users' Group hosted by the Cambridge Assessment Network. Neil Jones of ESOL presented a paper on 'Continuous calibration: an operational model for testing with multiple versions and sessions'.

### British Psychological Research Conference

Beth Black attended the British Psychological Research Conference in York in March. The programme featured keynotes and symposia involving internationally recognised scholars and specialist workshops to develop research skills.

### 6th International Conference on Multilevel Analysis

In May Carmen Vidal Rodeiro attended the 6th International Conference on Multilevel Analysis and presented a paper on 'The use of prior or concurrent measures of educational attainment when studying comparability of examinations using multilevel models'.

### Cambridge Assessment Conference

The third Cambridge Assessment Conference will take place on 15 October, 2007 at Robinson College, Cambridge. The theme of this year's conference will be the use of e-assessment and the likely impact that it will have on education. The keynote speaker will be Andrew Pollard of the ESRC Teaching and Learning Programme. The fee is £180 per delegate. For further information please email thenetwork@cambridgeassessment.org.uk or phone +44 (0) 1223 552830.

## Publication

An article by Martin Johnson, 'A review of vocational research in the UK 2002–2006', was published in the December issue of *The International Journal of Training Research* (Vol. 4, No. 2).

## The Psychometrics Centre

The Psychometrics Centre has appointed Professor Robert J. Sternberg and Professor James R. Flynn as Distinguished Associates. These prestigious professors will advise the Centre on specific research and applied activities, as well as on its overall strategic direction.

Robert J. Sternberg, who was the keynote speaker at last year's Cambridge Assessment Conference in October 2006, is Professor of Psychology and Director of the PACE (Psychology of Abilities, Competencies and Expertise) Centre at Tufts University (Massachusetts). His work at the PACE Centre is dedicated to the advancement of theory, research, practice and policy advancing the notion of intelligence as modifiable and capable of development throughout the life span.

James R. Flynn is Professor Emeritus at the University of Otago (New Zealand) and recipient of the University's Gold Medal for Distinguished Career Research. As a psychologist, he is best known for the 'Flynn effect', the discovery of massive IQ gains from one generation to another.

## Cambridge Assessment Network

### Certificate in the Principles and Practice of Assessment

This innovative course, offered in conjunction with the University of Cambridge, Institute of Continuing Education, provides a flexible approach to learning and is intended for all those in education and training interested in assessment issues, including teachers, examiners, exams officers, parents and employers. The course consists of three taught modules and a fourth module based on a personal study project. Each module is offered through weekly face-to-face tuition and online learning. A typical module lasts 10 weeks and the course:

- provides a grounding in the principles and practice of assessment;
- recognises participants' competence and work-place experience in relation to assessment, where applicable;
- offers opportunities for further personal and professional development, and career enhancement.

Each module is worth 15 credits and participants may choose to do any or all of the four modules. Successful completion of all four modules (60 credits) leads to the award of the Certificate of Continuing Education (Principles and Practice of Assessment) from the University of Cambridge Institute of Continuing Education. The course runs in Cambridge and Coventry. New modules begin in January and the fee is £400 per module. For further information please contact Dr Liz Morfoot (Certificate Programme Manager) on 01954 240280, email: certificate@cont-ed.cam.ac.uk

Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU

Tel: 01223 553854
Fax: 01223 552700
Email: ResearchProgrammes@cambridgeassessment.org.uk
http://www.cambridgeassessment.org.uk