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...’ 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. 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, Rodrigo 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 Irena Suto and Rita Nadas

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 focusing 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 Klobota, 1998; Royal-Wilson, 2005), personality (Braet et al., 1981; Creareox 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 Creatorex, 2006, in press). Task demands, on the other hand, might be influenced by a question’s length and features, the complexity and unfamiliarity 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 Kuffman, 1966; Raikes, 2007). A lot of these factors are probably 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: PA (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 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 Difference (MAD): 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 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.
The aim of our second study, which followed on directly from the first, was to identify question features that distinguish questions that are
the PE’s personal
Aim
Marking Expertise Project 1: Study 2

• Writing:
• whether single letter answers are required.
• the presence and importance of a diagram.

For physics it was concluded that seven features may be useful in
factors interact with one another to affect marking accuracy, could
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 fluidity: 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.

Conclusions

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 GCSE 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 GCSE 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).

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


References


Statements from the researchers:

1. Can we predict which questions will be marked accurately?
2. Can we predict which questions will be marked inaccurately?
3. Can we predict which questions will be marked accurately and which will be marked inaccurately?
4. Can we predict which questions will be marked accurately and which will be marked inaccurately, and why?

We are seeking to address these issues in our second Marking Expertise project, which focuses on GCSE 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 GCSE 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).

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 GCSE 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 GCSE 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).

Conclusions

We weighed the factors and concluded that accuracy is affected by expertise, experience, and knowledge. However, this would need to be handled very cautiously: effects on validity would need to be considered.


References


Statements from the researchers:

1. Can we predict which questions will be marked accurately?
2. Can we predict which questions will be marked inaccurately?
3. Can we predict which questions will be marked accurately and which will be marked inaccurately?
4. Can we predict which questions will be marked accurately and which will be marked inaccurately, and why?

Did examiners’ marking strategies change as they marked more scripts?

Dr Jackie Greatorex Research Division

Introduction
Previously (Suto and Greatorex, in press) it was thought 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 marking 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 GCSE-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, A-Level and International GCSEs (IGCSEs). 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.1) 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, effortless 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’ utilizes 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 recognize 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 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 Q1a 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 (PEx) wrote the question papers and led the marking. In larger examining teams the PExes ensured that the Team Leaders (TLeS) 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 marking examinations did some marking practice, 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 IGCSE and UK A-Level subjects were being used to mark IGCSE Mathematics and Business Studies examinations 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

<table>
<thead>
<tr>
<th>Examination</th>
<th>Examination Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology</td>
<td>Yan, Tc, 3a</td>
</tr>
<tr>
<td>Business Studies</td>
<td>1b, 1c, 4</td>
</tr>
<tr>
<td>English as a Second Language</td>
<td>1, 5, 7</td>
</tr>
<tr>
<td>French</td>
<td>1, 26, 37</td>
</tr>
<tr>
<td>Mathematics</td>
<td>1, 11, 21b</td>
</tr>
</tbody>
</table>

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 before the co-ordination meeting. 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 IGCSE and UK A-Level subjects were being used to mark IGCSE Mathematics and Business Studies examinations 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

<table>
<thead>
<tr>
<th>Examination</th>
<th>Examination Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology</td>
<td>Yan, Tc, 3a</td>
</tr>
<tr>
<td>Business Studies</td>
<td>1b, 1c, 4</td>
</tr>
<tr>
<td>English as a Second Language</td>
<td>1, 5, 7</td>
</tr>
<tr>
<td>French</td>
<td>1, 26, 37</td>
</tr>
<tr>
<td>Mathematics</td>
<td>1, 11, 21b</td>
</tr>
</tbody>
</table>

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
Biography 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 'scanning', from which it can be inferred that the 'scanning' strategy was used more 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', then 'no response', then 'evaluating', 'scanning' and 'no response'. Regarding differences in ratings between consecutive occasions there were no considerable differences. There was a noticeable difference in the ratings about 'scanning', which implies that the 'scanning' 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 'matching' 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 the 'scrutinising' strategy 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 'evaluating' strategy was used more during batch 1 than afterwards, and more before the co-ordination meeting than during the co-ordination sample;
- the 'scrutinising' strategy was used more during the co-ordination meeting than during the co-ordination sample, and more during the co-ordination sample than during batch 1.
In previous research we found that all the marking strategies were used in general GCE Business Studies and GCSE Mathematics, as well as Physics A-Level (Creatore 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 Creatore (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 Creatore (in press) predicted that examiners might use less ‘evaluating’ and ‘scrutinising’ 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, or 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. Notice 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 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 Creatore (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 in general GCE Business Studies and GCSE Mathematics, as well as Physics A-Level (Creatore 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 Creatore (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 Creatore (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 Creatore (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, or 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. Notice 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.
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 a evaluation process. 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, D or D would be best for the company. There were eight 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.

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

In 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

- "T" to indicate "never"
- "F" to indicate "frequently"
- "I" to indicate "intermittently"
- "A" to indicate "always"

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

- "T" to indicate "never"
- "F" to indicate "frequently"
- "I" to indicate "intermittently"
- "A" to indicate "always"

Questionnaire – Process of marking – French

Table 1

<table>
<thead>
<tr>
<th>Question 1</th>
<th>Before the co-ordination meeting</th>
<th>During the co-ordination process</th>
<th>During feedback 1</th>
<th>After feedback 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘matching’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘scanning’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘evaluating’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘no response’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Questionnaire – Process of marking – French

Table 2

<table>
<thead>
<tr>
<th>Question 20</th>
<th>Before the co-ordination meeting</th>
<th>During the co-ordination process</th>
<th>During feedback 1</th>
<th>After feedback 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘matching’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘scanning’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘evaluating’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘no response’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Questionnaire – Process of marking – French

Table 3

<table>
<thead>
<tr>
<th>Question 31</th>
<th>Before the co-ordination meeting</th>
<th>During the co-ordination process</th>
<th>During feedback 1</th>
<th>After feedback 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘matching’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘scanning’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘evaluating’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘no response’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

Table 4

<table>
<thead>
<tr>
<th>Whole examination paper</th>
<th>Before the co-ordination meeting</th>
<th>During the co-ordination process</th>
<th>During feedback 1</th>
<th>After feedback 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘matching’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘scanning’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘evaluating’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>‘no response’</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Acknowledgements

Jane Fuller 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


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-anchored 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 Milanova, 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 move to on-screen marking.

The research summarised here started by reviewing the relevant literature in the area 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), Pala and Huht (1993) and Huht (1993) in the context of assessing writing. Milanova, 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 A-level and an A2 level geography exam paper were selected. The A2 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 for each examiner marking 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...
ISSUE 4

RESEARCH MATTERS

Marking and marking behaviours

Marking was broadly in line with marking but that examiners tended towards severity in comparison. One examiner's marking of the A2 unit was more severe than the other and out of line with line 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 about the small number of repeated scripts suggested that thinking about the marking required attention, but.

Coding the verbal protocols

Scripts of the verbal protocols were analysed to try to understand the processes involved in the marking. Drawing on the transcripts and the theory of Sanderson (2001) and Milovac 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 coders were grouped into the categories of:

- "reading and understanding" (codes relating to reading and making sense of responses);
- "evaluation" (codes relating to evaluating a response or part of a response);
- "language" (codes relating to the student's use of language);
- "personal perception" (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 in terms of length of response, addressing/understanding questions);
- "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 unexpectedly) were behaviours relating to evaluating the text being read. Clearly positive and negative evaluations (e.g. "so that's a good evaluation point", "so 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 equally prominent.

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 A2 exam 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. "quite like that", "London [green] my heart drop", 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 this 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 so that's a good evaluation point", "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.

<table>
<thead>
<tr>
<th>Transcript extract</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Now we have got Mexico, Mexico city from rural areas, ok.</td>
<td>Summaries/paraphrases positive evaluation</td>
</tr>
<tr>
<td>reasoning at a rate, mention job opportunities, well explained them.</td>
<td>Summaries/paraphrases positive evaluation</td>
</tr>
<tr>
<td>a high standard, cramped housing, talking about it is like in Mexico city rather than the whole ( ).</td>
<td>Summaries/paraphrases neutral evaluation</td>
</tr>
<tr>
<td>shanty towns, now it's gone to a real town.</td>
<td>Summaries/paraphrases Negative evaluation and relevance</td>
</tr>
<tr>
<td>It's been a bit of a task and explanation in only one area, (using mark scheme) ()</td>
<td>Reference to mark scheme</td>
</tr>
<tr>
<td>as it's level 2 it's fairly general</td>
<td>First indication of mark</td>
</tr>
<tr>
<td>so I think we will give that 5</td>
<td>Mark decision</td>
</tr>
<tr>
<td>because it hasn't really explained much more than, not a lot about what it is in the area they've come from, so really only explaining one area, somehow</td>
<td>Discussion/review of mark/ re-assessment</td>
</tr>
</tbody>
</table>

Findings

Did the frequencies of coded behaviours and reactions vary between the marking of different types of questions (short and medium length 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 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 examers (one of 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 examiner 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 examiner 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 topical question. 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-occurring 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 ‘prolegomenon’ and ‘epilogue’ phases before reading begins and after mark decisions have been made. The model attempts to...

Discuss the findings of the research in the context of its implications for the marking process and educational practices. Consider the role of marking in shaping student outcomes and the potential for developing more effective and fair assessment methods. Reflect on the importance of understanding the cognitive processes involved in marking and how these can be improved to enhance educational standards and student performance.

**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 'headings' and 'summarises/paraphrases' were among the most frequently applicable codes and the frequency of reading behaviours seemed to be associated with marker agreement.

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 Creemers’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, highlighting 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 aspirers towards criterion-referenced standards (Baird, Crosswell & 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...
Quality control of examination marking

John F. Bell, Tom Bramley, Mark J. A. Claassen 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 and 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 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 the minimum amount of marking that must be reviewed or re-done.

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 of the candidates' marking.
- The amount of computationally intensive, iterative processing needed.
- The difficulty of explaining the model to stakeholders with little or no technical knowledge.
- The amount of knowledge required.
- 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 Rasch models.
- The standard deviation of a normal distribution with mean zero from which an error factor for each marker on that item will be drawn at random for each error factor.
- The means model.
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 doing, 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 are around 0.2 logits, and none 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 highest 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.

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 logits 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.

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


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 Bondombo, 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 completely 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, low1 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 objective1 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 ‘read’ into a marker’s allocation of scripts to be marked. I have arbitrarily set the mark limit for this scenario to 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 fail 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 x m x n x m x 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’.1

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

<table>
<thead>
<tr>
<th>Correct mark</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Row sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>14</td>
<td>1</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

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 (PR) which is the proportion of the total frequency coming from the shaded cells.

\[ PR = \frac{\sum n_{ij}}{N} \]

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 ways statistical position to measuring agreement with dichotomous (1–3 marks) 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 are considerable difficulties in implementing a computer algorithm for marking ‘fully 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 x n x m x m x m table 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

<table>
<thead>
<tr>
<th>Correct mark</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Row sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>14</td>
<td>1</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

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 3: Percentages of differences between observed and correct mark

<table>
<thead>
<tr>
<th>Correct mark</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Row sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>14</td>
<td>1</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

In this form, the table shows the percentage of cases where the observed and correct mark were the same, the percentage of cases where the observed mark was higher or lower than the correct mark, and the percentage of cases where the observed mark was more than one, two or three marks away from the correct mark.

This might be extended to include, for example, an examination mark scheme (in Table 4 m = 3) where the correct mark is taken as the PE’s mark, and for comparison purposes it might be helpful to show frequencies rather than percentages in the table, as in Table 4 below.

Table 4: Percentages of differences between observed and correct mark

<table>
<thead>
<tr>
<th>Item max</th>
<th>Difference</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>-3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>90</td>
<td>-3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>80</td>
<td>-3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>70</td>
<td>-3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>60</td>
<td>-3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>50</td>
<td>-3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>40</td>
<td>-3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>30</td>
<td>-3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>20</td>
<td>-3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>-3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

In this form, the table shows the percentage of cases where the observed mark was the same as the correct mark, the percentage of cases where the observed mark was higher or lower than the correct mark, and the percentage of cases where the observed mark was more than one, two or three marks away from the correct mark.

In practice there are considerable difficulties in implementing a computer algorithm for marking ‘fully objective questions’ – see, for example, Sukkarieh et al (2003).
marks awarded in each level. If the mark scheme is levels-based, there may be a natural ‘bin’ corresponding to the range of responses on the trait, whereas erraticism is differences by considering the latter as differences 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.

The broad lines superimposed on the plot show the mean difference, its standard error and the SD (or IQR) of the differences indicates the extent to which the marker’s interpretation of the ‘true’ 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 means that (even more than with scenario 1) a single statistic cannot capture all the relevant information about marker agreement. 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 (P0) was relatively low. This might happen, for example, if the observed mark was consistently one mark higher than the correct mark. 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, as shown in Table 6 below (which uses the same percentages as Table 4).

3 Conceptually, erraticism can be distinguished from differences in interpretation of the ‘true’ trait by considering the latter as differences between marks in terms of which position the response to lie on the trait, whereas erraticism is differences within a marker as to where they perceive the same response to lie or hypothetical implications of marking in practice, these two are difficult to separate.

4 If the rank scheme is levels-based, there may be a natural ‘bin’ corresponding to the range of marks awarded in each level.
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. For the results to apply more generally the response 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, $\rho_{agr}$ has been suggested, which is the proportion of scripts with a difference between marker and PE in a ±SD range around zero. $N$ could be increased as the total mark for the question (or sub-test or test) increases.

For full discussion 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 likely 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 rather how ‘reliable’ it is 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 covariation estimates the reliability. Alpha is referred to as an ‘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 critical 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 Eyn and Myn, 2005). Some see this as an advantage, others as a disadvantage (Uebersax, 2003). Most versions of the ICC require double or multiple sets 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 multilevel 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. 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 $\pm z \times SEM$. 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.

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 Steiner (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 Wolle (2003, 2004) show that it is possible to use the output from a many-facet Rasch analysis to diagnose other rater errors such as central tendency (overusing the middle categories of the mark scale), a halo effect (tendency to award similar marks to the same candidate on different questions) and differential severity/leniency (a tendency to be more lenient towards particular subsets of candidate).

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 split is 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 Hawton (2005) and risk the short-term cost in terms of public trust by becoming maker to report indices of marker agreement. If they do, it will be important to choose indices which reveal more than 6. This will also help to mitigate any floor and ceiling effects when comparing differences between marker and PE.

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.
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

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:

- Discuss and negotiate the marks on all the differences or on specified differences.
- 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 Foukou (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.
- Resort to a third marker, who could mark the script afresh or, based on the previous two 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 ‘P’ (Vere 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.
2. Re-marking of the same 200 scripts using three different strategies:
   - Treatment 1: Blind re-marking by two assistant examiners (marking 100 scripts each) plus the reconciliation by the PE as needed.
   - Treatment 2: Blind re-marking by two senior (or experienced) examiners (marking 100 scripts each) plus the reconciliation by the PE as needed.
   - 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 \( r \) 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 focused 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:

\[
\kappa = \frac{P_o - P_e}{1 - P_e},
\]

where \( P_o \) is the proportion of marks in which the markers agree and \( P_e \) is the proportion of marks for which agreement is expected by chance.

Table 1 provides evidence that agreement is highest 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.

### 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.

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

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>200</td>
<td>43.72</td>
<td>9.06</td>
<td>17</td>
</tr>
<tr>
<td>Treatment 1</td>
<td>200</td>
<td>44.05</td>
<td>8.82</td>
<td>17</td>
</tr>
<tr>
<td>Treatment 2</td>
<td>200</td>
<td>43.93</td>
<td>9.15</td>
<td>15</td>
</tr>
<tr>
<td>Treatment 3</td>
<td>200</td>
<td>44.09</td>
<td>9.19</td>
<td>18</td>
</tr>
</tbody>
</table>

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 increases. For example, for treatment 1 (blind re-marking by assistant examiner) the marks differ by no more than +/- one in around 43% of the scripts marked and by +/- three in around 78% 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.

### Correlation coefficients

The correlation coefficients are high (the smallest correlation appears between the original mark and treatment 1: \( r = 0.93538 \)) 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.

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>
<thead>
<tr>
<th>Treatment</th>
<th>Blind</th>
<th>Assistant</th>
<th>Blind</th>
<th>Non-blind</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>200</td>
<td>43.72</td>
<td>9.06</td>
<td>17</td>
</tr>
<tr>
<td>Treatment 1</td>
<td>200</td>
<td>44.05</td>
<td>8.82</td>
<td>17</td>
</tr>
<tr>
<td>Treatment 2</td>
<td>200</td>
<td>43.93</td>
<td>9.15</td>
<td>15</td>
</tr>
<tr>
<td>Treatment 3</td>
<td>200</td>
<td>44.09</td>
<td>9.19</td>
<td>18</td>
</tr>
</tbody>
</table>

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 a 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.

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.

### 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 subjective judgement by markers.

Table 10 : Absolute differences in marks

<table>
<thead>
<tr>
<th>Difference in marks</th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; -13</td>
<td>1.0</td>
<td>4.5</td>
<td>0.0</td>
<td>1.8</td>
</tr>
<tr>
<td>-13 to -11</td>
<td>1.5</td>
<td>1.5</td>
<td>0.0</td>
<td>2.3</td>
</tr>
<tr>
<td>-10 to -8</td>
<td>6.5</td>
<td>15.0</td>
<td>2.5</td>
<td>8.0</td>
</tr>
<tr>
<td>-7 to -5</td>
<td>8.0</td>
<td>15.0</td>
<td>5.0</td>
<td>39.3</td>
</tr>
<tr>
<td>-4 to -2</td>
<td>10.5</td>
<td>23.0</td>
<td>20.1</td>
<td>57.7</td>
</tr>
<tr>
<td>-1</td>
<td>7.5</td>
<td>5.5</td>
<td>15.5</td>
<td>9.5</td>
</tr>
<tr>
<td>0</td>
<td>8.0</td>
<td>2.5</td>
<td>35.5</td>
<td>15.3</td>
</tr>
<tr>
<td>3</td>
<td>6.5</td>
<td>8.5</td>
<td>7.5</td>
<td>7.5</td>
</tr>
<tr>
<td>2 to 4</td>
<td>10.5</td>
<td>11.1</td>
<td>9.5</td>
<td>13.4</td>
</tr>
<tr>
<td>4 to 6</td>
<td>10.0</td>
<td>6.5</td>
<td>3.0</td>
<td>6.5</td>
</tr>
<tr>
<td>6</td>
<td>6.0</td>
<td>4.0</td>
<td>1.0</td>
<td>7.7</td>
</tr>
<tr>
<td>&gt; 10</td>
<td>9.0</td>
<td>1.0</td>
<td>0.0</td>
<td>7.0</td>
</tr>
</tbody>
</table>

Oxidised silver and non-blinded re-marked scripts did not have to be reconciled. The former was slightly higher in treatment 3 (non-blinded re-mark), the mean was 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.

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.

### 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 subjective judgement by markers.

Table 11 : Distribution of differences between original and experimental marks

<table>
<thead>
<tr>
<th>Difference in marks</th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; -13</td>
<td>1.0</td>
<td>4.5</td>
<td>0.0</td>
<td>1.8</td>
</tr>
<tr>
<td>-13 to -11</td>
<td>1.5</td>
<td>1.5</td>
<td>0.0</td>
<td>2.3</td>
</tr>
<tr>
<td>-10 to -8</td>
<td>6.5</td>
<td>15.0</td>
<td>2.5</td>
<td>8.0</td>
</tr>
<tr>
<td>-7 to -5</td>
<td>8.0</td>
<td>15.0</td>
<td>5.0</td>
<td>39.3</td>
</tr>
<tr>
<td>-4 to -2</td>
<td>10.5</td>
<td>23.0</td>
<td>20.1</td>
<td>57.7</td>
</tr>
<tr>
<td>-1</td>
<td>7.5</td>
<td>5.5</td>
<td>15.5</td>
<td>9.5</td>
</tr>
<tr>
<td>0</td>
<td>8.0</td>
<td>2.5</td>
<td>35.5</td>
<td>15.3</td>
</tr>
<tr>
<td>3</td>
<td>6.5</td>
<td>8.5</td>
<td>7.5</td>
<td>7.5</td>
</tr>
<tr>
<td>2 to 4</td>
<td>10.5</td>
<td>11.1</td>
<td>9.5</td>
<td>13.4</td>
</tr>
<tr>
<td>4 to 6</td>
<td>10.0</td>
<td>6.5</td>
<td>3.0</td>
<td>6.5</td>
</tr>
<tr>
<td>6</td>
<td>6.0</td>
<td>4.0</td>
<td>1.0</td>
<td>7.7</td>
</tr>
<tr>
<td>&gt; 10</td>
<td>9.0</td>
<td>1.0</td>
<td>0.0</td>
<td>7.0</td>
</tr>
</tbody>
</table>

Oxidised silver and non-blinded re-marked scripts did not have to be reconciled. The former was slightly higher in treatment 3 (non-blinded re-mark), the mean was 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.

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.

### 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 subjective judgement by markers.

Table 12 : Pearson’s correlation coefficients

| Original – Treatment 1 | 0.6951 | 0.6789 | 0.7276 | 0.7417 |
| Original – Treatment 2 | 0.6593 | 0.6789 | 0.7276 | 0.7417 |
| Original – Treatment 3 | 0.9464 | 0.7417 | 0.7276 | 0.7417 |

In English, the number of scripts needing reconciliation was much higher 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.

### 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 subjective judgement by markers.

Table 13 : Scripts that needed reconciliation

<table>
<thead>
<tr>
<th>Reconciliation</th>
<th>Original – Treatment 1</th>
<th>Original – Treatment 2</th>
<th>Original – Treatment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind</td>
<td>780</td>
<td>780</td>
<td>780</td>
</tr>
<tr>
<td>Blind/Non-blind</td>
<td>780</td>
<td>780</td>
<td>780</td>
</tr>
</tbody>
</table>

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.

### 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 subjective judgement by markers.

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

<table>
<thead>
<tr>
<th>Difference in marks</th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; -10</td>
<td>1.0</td>
<td>4.5</td>
<td>12.0</td>
</tr>
<tr>
<td>-10 to -8</td>
<td>6.5</td>
<td>15.0</td>
<td>23.0</td>
</tr>
<tr>
<td>-8 to -5</td>
<td>8.0</td>
<td>15.0</td>
<td>31.0</td>
</tr>
<tr>
<td>-5 to -3</td>
<td>10.5</td>
<td>23.0</td>
<td>41.0</td>
</tr>
<tr>
<td>-3 to -1</td>
<td>12.0</td>
<td>23.0</td>
<td>51.0</td>
</tr>
<tr>
<td>-1 to 0</td>
<td>12.0</td>
<td>23.0</td>
<td>61.0</td>
</tr>
<tr>
<td>0</td>
<td>12.0</td>
<td>23.0</td>
<td>71.0</td>
</tr>
<tr>
<td>1</td>
<td>12.0</td>
<td>23.0</td>
<td>81.0</td>
</tr>
<tr>
<td>&gt; 10</td>
<td>12.0</td>
<td>23.0</td>
<td>91.0</td>
</tr>
</tbody>
</table>

Oxidised silver and non-blinded re-marked scripts did not have to be reconciled. The former was slightly higher in treatment 3 (non-blinded re-mark), the mean was 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.

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.

### 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 subjective judgement by markers.
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

Abstract
Studies of inter-marker reliability in GCSE and A-level examinations have been reported in the literature, but typically those focused on paper totals, rather than item marks. See, for example, Newton (1995). 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 GCSE A-level and GCSE 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-marker 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-marker reliability, but typically these focused on agreement at the level of the total mark given to scripts. The operational procedures followed by examination Boards for documentation 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-marker 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 marking study was low stakes for the examiners (i.e. no candidate’s result depends 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:
- **ICCS Foreign Language French:**
  - Multiple choice (m/c) and short answer textual answers worth 1–2 marks
  - ICCS Development Studies: Alternative to Coursework
  - Short answers worth 1–6 marks
  - A-level Chemistry: Structured Questions
  - A-level Mathematics: 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-marker 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. 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’ 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 Research Matters, but correlation based measures are useful when considering the relationship between more than two examiners, as is the case here.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Max mark</th>
<th>Min mark</th>
<th>Time (mins) per mark</th>
<th>ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>45</td>
<td>40</td>
<td>0.9</td>
<td>0.917</td>
</tr>
<tr>
<td>Dev. Stud.</td>
<td>35</td>
<td>30</td>
<td>2.6</td>
<td>0.922</td>
</tr>
<tr>
<td>Chemistry</td>
<td>60</td>
<td>50</td>
<td>1.0</td>
<td>0.922</td>
</tr>
<tr>
<td>Economics</td>
<td>110</td>
<td>90</td>
<td>3.1</td>
<td>0.774</td>
</tr>
<tr>
<td>Sociology</td>
<td>50</td>
<td>40</td>
<td>1.8</td>
<td>0.685</td>
</tr>
</tbody>
</table>

*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.

At 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 0.98 – 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-marker 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.
- **Points based marking** – these items generally require brief responses and greatly constrain how candidates must respond. Examples might include marking to the test: a framework to assess if it is worth the trouble.
- **Levels based marking** – these items generally require longer responses and may also require the marker to apply a principle of best fit when deciding the mark for a response.

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

ASSURING QUALITY IN ASSESSMENT

Item-level examiner agreement
Nicholas Raikes and Alf Massey
Research Division

Table 1: Intraclass correlations for script totals

Inter-examiner agreement at script-total level

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.
- **Points based marking** – these items generally require brief responses and greatly constrain how candidates must respond. Examples might include marking to the test: a framework to assess if it is worth the trouble.
- **Levels based marking** – these items generally require longer responses and may also require the marker to apply a principle of best fit when deciding the mark for a response.
The maximum mark available for the examination, i.e. it is the average time available to candidates for earning a mark.

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.

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 more reliably marked 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 using a levels marking scheme were marked very reliably, proving that it is possible to achieve high reliability for essay marking.

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 more reliably marked 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 using a levels marking 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 that was imposed on candidates and the intraclass correlation (ICC) obtained for script total marks. A Pearson correlation of −0.59 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."

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

<table>
<thead>
<tr>
<th>Max mark</th>
<th>Dev. Stud.</th>
<th>Chemistry</th>
<th>Economics</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.975</td>
<td>0.981</td>
<td>0.950</td>
<td>0.978</td>
</tr>
<tr>
<td>2</td>
<td>0.986</td>
<td>-</td>
<td>-</td>
<td>0.978</td>
</tr>
<tr>
<td>6</td>
<td>0.975</td>
<td>0.980</td>
<td>0.950</td>
<td>0.978</td>
</tr>
<tr>
<td>All</td>
<td>0.975</td>
<td>0.980</td>
<td>0.950</td>
<td>0.978</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean ICC</th>
<th>Standard Deviation of the ICCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Nitems)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.975</td>
</tr>
<tr>
<td>2</td>
<td>0.986</td>
</tr>
<tr>
<td>6</td>
<td>0.975</td>
</tr>
<tr>
<td>All</td>
<td>0.975</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Max mark</th>
<th>Dev. Stud.</th>
<th>Chemistry</th>
<th>Economics</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.950</td>
<td>-</td>
<td>-</td>
<td>0.978</td>
</tr>
<tr>
<td>2</td>
<td>0.978</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>0.950</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>All</td>
<td>0.950</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean ICC</th>
<th>Standard Deviation of the ICCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Nitems)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.950</td>
</tr>
<tr>
<td>2</td>
<td>0.978</td>
</tr>
<tr>
<td>6</td>
<td>0.950</td>
</tr>
<tr>
<td>All</td>
<td>0.950</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Mean ICC</th>
<th>Standard Deviation of the ICCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Nitems)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.990</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
</tr>
<tr>
<td>25</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Max mark</th>
<th>Dev. Stud.</th>
<th>Chemistry</th>
<th>Economics</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.990</td>
<td>-</td>
<td>-</td>
<td>0.990</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>0.740</td>
<td>-</td>
<td>0.740</td>
</tr>
<tr>
<td>10</td>
<td>-</td>
<td>0.567</td>
<td>-</td>
<td>0.567</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>0.585</td>
<td>-</td>
<td>0.585</td>
</tr>
<tr>
<td>25</td>
<td>-</td>
<td>0.825</td>
<td>0.825</td>
<td>0.890</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean ICC</th>
<th>Standard Deviation of the ICCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Nitems)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.990</td>
</tr>
<tr>
<td>8</td>
<td>0.740</td>
</tr>
<tr>
<td>10</td>
<td>0.567</td>
</tr>
<tr>
<td>12</td>
<td>0.585</td>
</tr>
<tr>
<td>25</td>
<td>0.825</td>
</tr>
</tbody>
</table>

| All      | 0.990                           |
|          | 0.631                           |
|          | 0.825                           |
|          | 0.773                           |

<table>
<thead>
<tr>
<th>Mean ICC</th>
<th>Standard Deviation of the ICCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Nitems)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.990</td>
</tr>
<tr>
<td>8</td>
<td>0.740</td>
</tr>
<tr>
<td>10</td>
<td>0.567</td>
</tr>
<tr>
<td>12</td>
<td>0.585</td>
</tr>
<tr>
<td>25</td>
<td>0.825</td>
</tr>
</tbody>
</table>

| All      | 0.990                           |
|          | 0.631                           |
|          | 0.825                           |
|          | 0.773                           |

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 a judgemental system 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 a judgemental system that was imposed on candidates and 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 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 item). It is not 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.

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 be 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?

References


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 ongoing 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 ‘free’ 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:‘[S]uch 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


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. 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 Group hosted by the Cambridge Assessment Network. Neil Jones of ESOR, 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 keynote 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 thework/Cambridgeassessment.org.uk or phone +44 (0) 1223 552830.

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, exam 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

The Psychometrics Centre

The Psychometrics Centre has appointed Professor Robert J. Stemberg 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 overall strategic direction.

Robert J. Stemberg, 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.

Research News

What can we measure? And what should we be measuring?


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 ongoing 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 ‘free’ 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:‘[S]uch 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


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. 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 Group hosted by the Cambridge Assessment Network. Neil Jones of ESOR, 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 keynote 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 thework/Cambridgeassessment.org.uk or phone +44 (0) 1223 552830.

Publication
