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

Dr Jackie Greatorex Research Division

Introduction

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

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

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

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

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

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

For these particular IGCSEs the Principal Examiners (PEs) wrote the question papers and led the marking. In larger examining teams the PEs ensured that the Team Leaders (TLs) were marking to the same standard as the Principal Examiner. The Team Leaders ensured the quality of the

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

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

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

Method

Questionnaire development

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

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

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

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

Table 1 : The examination questions included in the questionnaire

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

Administration

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

Participants

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

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

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

Results

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

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

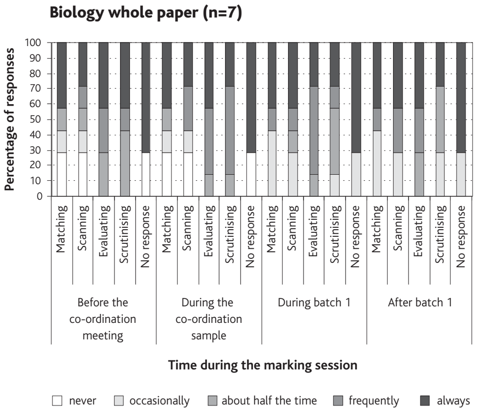


Figure 1 : Graph to show the percentage of ratings for Biology examiners

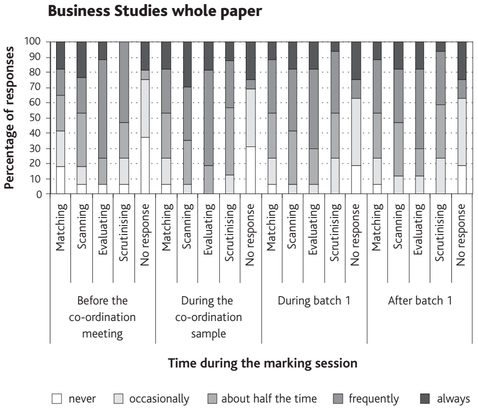


Figure 2 : Graph to show the percentage of ratings for Business Studies examiners

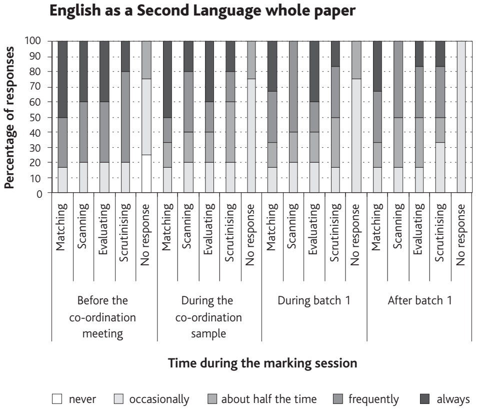


Figure 3 : Graph to show the percentage of ratings for English as a Second Language examiners

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

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

Biology whole examination

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

Business Studies whole examination

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

English as a Second Language whole examination

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

- the 'no response' strategy was used more during the co-ordination sample than before the co-ordination meeting;

- the ‘evaluating’ strategy was used more during batch 1 than afterwards, and more before the co-ordination meeting than during the co-ordination sample;
- the ‘scanning’ strategy was used more before the co-ordination meeting than during the co-ordination sample, and more during the co-ordination sample than during batch 1.

French whole examination

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

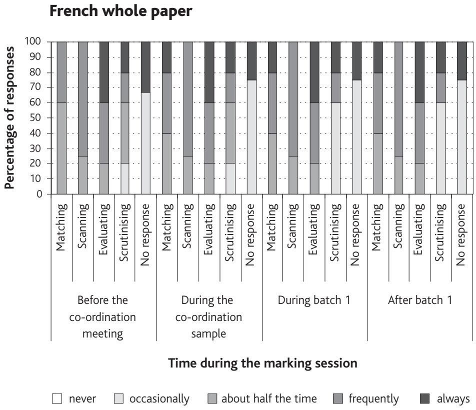
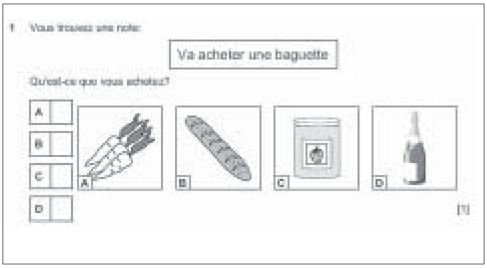


Figure 4: Graph to show the percentage of ratings for French examiners


1 Vous trouvez une note:

Va acheter une baguette

Qu'est-ce que vous achetez?

A

B

C

D

[1]

French Question Paper extract

1 B **1 Mark per item**

Mark Scheme extract

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

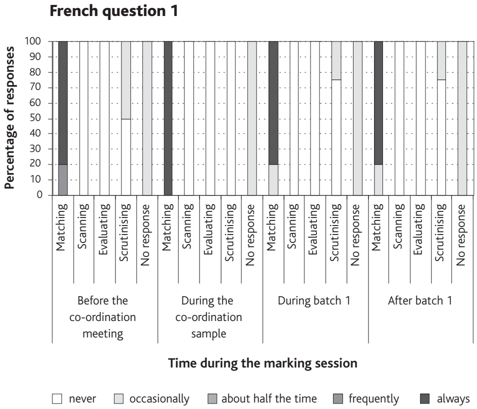
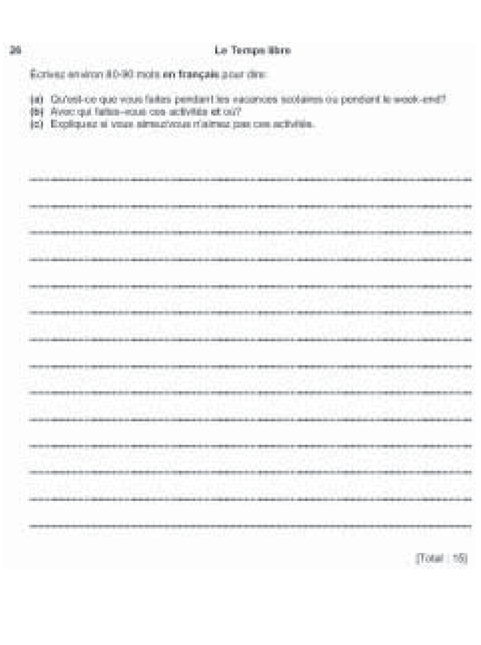


Figure 5 : Graph to show the percentage of ratings for French examiners

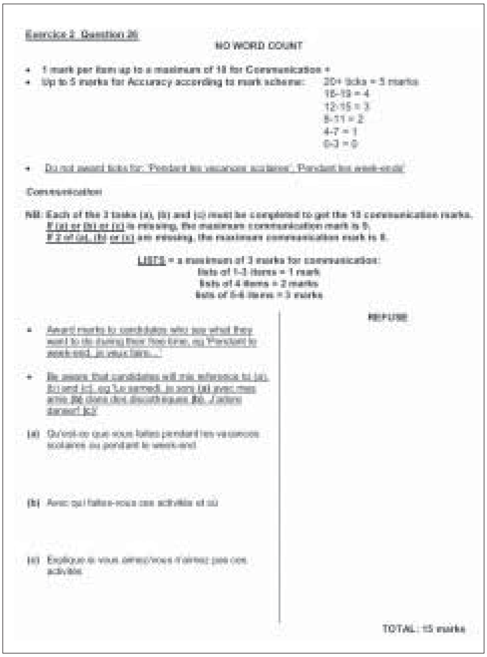

26 Le Temps libre

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

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

[Total : 15]

French Question Paper extract



Exercice 2 Question 26

NO WORD COUNT

- 1 mark per item up to a maximum of 10 for Communication +
- Up to 5 marks for Accuracy according to mark scheme: 20+ ticks = 5 marks
  16-19 = 4
  12-15 = 3
  8-11 = 2
  4-7 = 1
  0-3 = 0

- Do not award ticks for: 'Pendant les vacances scolaires', 'Pendant les week-ends'

Communication

NB: Each of the 3 tasks (a), (b) and (c) must be completed to get the 10 communication marks.
If (a) or (b) or (c) is missing, the maximum communication mark is 8.
If 2 of (a), (b) or (c) are missing, the maximum communication mark is 5.

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

REFUSE

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

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

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

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

TOTAL: 15 marks

Mark Scheme extract

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

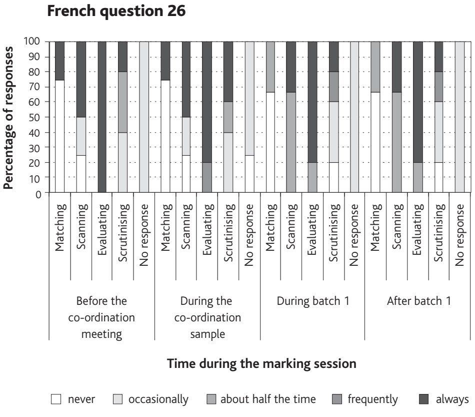


Figure 6 : Graph to show the percentage of ratings for French examiners

Mathematics whole examination

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

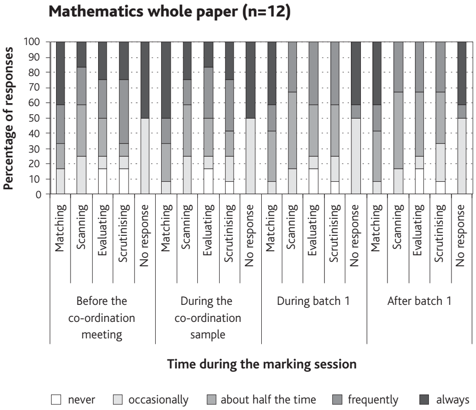


Figure 7 : Graph to show the percentage of ratings for Mathematics examiners

Conclusion and Discussion

The research was limited by some factors.

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

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

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

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

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

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

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

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

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

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

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

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

APPENDIX 1

Marking Strategies Reference Sheet (updated)

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

Matching

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

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

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

Scanning

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

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

Evaluating

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

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

Scrutinising

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

No response

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

APPENDIX 2

Questionnaire – Process of marking – French

INSTRUCTIONS

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

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

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

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

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

Acknowledgements

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

References

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

Greatorex, J. & Suto, W. M. I. (2006). *An empirical exploration of human judgement in the marking of school examinations*. A paper presented at the International Association of Educational Assessment conference, May 2006, Singapore.

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

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

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

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

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

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

Researching the judgement processes involved in A-level marking

Victoria Crisp Research Division

Introduction

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

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

Method

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

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

The scripts marked were photocopies of live scripts with marks and annotations removed. Examiners marked the same students' scripts, except that in the silent marking and think aloud marking, for each