

## USING THE THINK ALOUD METHOD TO INVESTIGATE ELECTRONIC MARKING OF LONG FORM ANSWERS

### ABSTRACT

This paper considers the extension of electronic marking beyond short written answers to 'Long Form Answers' (LFAs) which are typically, but not exclusively, responses to essay questions. Current written components have all been drawn into the AQA/DRS e-marking programme wherever they are deemed suitable. To date AQA has not felt confident in allocating components containing LFAs to e-marking, both for scanning reasons and, most importantly, because of lack of assurance that such components can be reliably marked on screen. An earlier investigation into marking GCSE English essays on screen had not given such assurance (Fowles, 2006). Therefore, before investigating the reliability of e-marking LFAs in other subject areas, it was decided to carry out a qualitative exploration of how examiners experience marking them on screen, checking for any suggestion of limits to e-marking long written responses or of interference with the cognitive processes involved in marking.

Two essay questions were selected, from a GCE Biology and a GCSE French paper, and a small number of examiners were invited to mark some examples of candidates' responses to these questions on screen using the annotation tool as currently developed in CMI+. The exercise used the 'think aloud' method, which meant recording all the examiners' comments and verbalised thoughts as they marked, to provide verbal data for analysis. The examiners were asked to talk aloud continuously as they marked on screen, about whatever they were thinking as they marked. They were aware that the focus of the research was on how they were experiencing marking LFAs on screen, in particular on whether the volume of written response could become too great for them to feel confident in the accuracy of their marking.

The recordings show that the examiners responded in individual ways to the request to think aloud while marking on screen, and their verbal records tended to be comments of two kinds. The first referred to interactions with CMI+, especially the annotation tool, which generated some useful suggestions for software improvements, such as alternatives ways to navigate around an LFA. The second were comments in justification of the marks being awarded. In the latter the Biology/French content of the scripts took over the examiners' expressed thoughts. While this is a positive outcome in that the transfer from paper to e-marking was apparently made swiftly and without need to comment, it did not support the use of the think aloud method alone to investigate issues such as limits, if any, to response length in e-marking LFAs. Alternative methods, both concurrent and retrospective, to combine with the think aloud method are discussed. They might be considered for use in the future, as AQA's planned e-marking programme transfers increasingly complex LFAs from paper-based to electronic marking, should any issues be identified with regard to marking quality.

*Key words: electronic marking, essay questions, long form answers, talk-aloud method*

## INTRODUCTION

### Background

AQA has commissioned a rapid expansion of electronic mark capture, especially electronic marking. Written components have all been drawn into the AQA/DRS CMI+ e-marking programme wherever they are deemed suitable. Unsuitable components tend to be those that present difficulties for scanning, such as reliance on colour in candidates' responses, responses not written in prescribed areas, unusual sized paper (e.g. A3 rather than A4), and choice of either whole sections or questions within sections. Choice and longer essay-style questions complicate scanning because the answers are usually written in free-form answer booklets instead of pre-defined locations or 'clip areas' in a question paper booklet. The introduction to e-marking of questions with high mark allocations and long written answers, labelled within DRS as 'Long Form Answers' (LFAs), has been deferred in large part because of these scanning difficulties.

In addition to the technical and practical difficulties of capturing LFAs for on-screen marking, most importantly there is also a lack of research evidence that marking reliability would not suffer in transferring them from paper to electronic marking. Research studies exploring the effectiveness of e-marking appear to be lacking in the literature and only a small number could be identified by Taylor (2007) in her report on interpreting Enquiries after Results (EAR) data as a measure of quality of marking. One study conducted by Fowles (2002) compared the conventional marks given to GCE Chemistry scripts from the January 2002 examination series with those given during on-screen marking of the same scripts following the examination period. A close relationship was found between the two sets of marks, and examiners were not found to mark the short answer questions on the paper any more severely on screen than on paper. Raikes (2002) compared the effectiveness of conventional paper marking with on-screen marking for three GCE components, in Mathematics, English Literature and Geography. Scripts were presented on screen for both whole paper and segmented, or question by question, marking. The findings were that the examiners in Mathematics were equally consistent across the three conditions, that in Geography, although the majority of marking was consistent, one examiner was found to mark more severely on screen than on paper and another more inconsistently on screen than on paper, while the examiners marking English Literature tended to be most consistent when marking on paper, and least so when marking on screen at question level. Raikes was able to conclude that on-screen marking is likely to be as reliable as paper-based marking, though it was suggested that segmentation would require further investigation, which still appears to be the case.

More recently, an evaluation for QCA was conducted by AQA of live CMI+ e-marking in January 2005, involving GCSE French Specification B listening tests (Fowles, 2005). Scripts were segmented and the questions identified as requiring human marking (i.e. expert or general marking) were double marked. The findings showed a high level of agreement between examiners for the majority of responses (98.4%). However, the question papers involved in this study, like the earlier GCE Chemistry paper marked on screen (Fowles, 2002), as well as those now e-marked in practice, tend to involve short response questions, meaning high levels of agreement between examiners is perhaps not unexpected. To date AQA has not felt confident in allocating components containing LFAs to e-marking through lack of assurance that such components can be reliably marked on screen. For this reason a further investigation of e-marking was carried out in 2006 involving GCSE English (Fowles, 2006b). English was chosen following an earlier investigation into the reliability of (paper-based) marking in the current AQA GCSE English specifications, which demonstrated a high level of reliability for the

Higher tier of Specification A in the 2005 examination (Fowles, 2006a). One of the two Specification A Higher tier papers was therefore selected to investigate e-marking, in the hope that it might guide the selection of components for e-marking in other subjects where the nature of candidates' responses has been regarded within AQA as challenging for marking using CMI+.

The addition of an annotation tool was introduced into CMI+ for a few components marked in the summer of 2006. The tool allows examiners to place marks and include comments on the screen image at appropriate points in the candidates' responses, imitating their paper-based marking practices. Electronic marking in the GCSE English trial therefore used this extra feature. The selected question paper was made up of five part questions in Section A and a choice of one from four essay questions in Section B. Candidates' responses were written in a separate answer booklet. The six examiners in the trial also re-marked a second sample of scripts on paper. The Principal Examiner was included as one of the six, and the purpose of marking the sample on paper was to establish the relationship of each examiner's marking to that of the Principal Examiner, to provide a context in which to evaluate the marking with CMI+. In both the paper and the electronic contexts, each examiner's marks on the two sets of sample scripts were compared with the Principal Examiner's 'true' marks. Thus the same assumption was made as in the earlier, paper-based investigation (Fowles, 2006a), that the Principal Examiner's marks stand as the 'true' marks. The comparison of marking in the e-marking trial used the same approaches as the earlier study and included calculating the coefficients of correlation between the true and the individual examiner's marks, a regularly used indicator of marking reliability. The results gave some suggestion that marking of the long essay question of Section B had not transferred well to CMI+. Specifically, the coefficients of correlation in the paper-based study had suggested that marker reliability was similar in both contexts for Section A and for Section B on paper, ranging from 0.60 to 0.78. However Section B appeared less reliably marked electronically, with the range of correlation falling to between 0.29 and 0.43 (Fowles, 2006b).

As already noted, electronic marking with CMI+ raises the issue of differences, if any, between whole paper and segmented marking. Some of the GCSE English examiners suggested that they normally keep an awareness, as they are marking, of the candidate's performance on the whole paper and they become 'familiar' with a candidate, including their handwriting, in Section A, which, it was claimed, helps them in reading and marking the essay in Section B. This cannot happen in electronic marking with CMI+, and therefore proponents of whole paper marking feel that it cannot give such accurate marking over the whole paper. However AQA continues to prefer segmented e-marking (within CMI+) to whole paper e-marking because it has greater potential to monitor examiners' marking on individual questions and to direct them to specialist questions, as well as being considered more objective and reliable. The impact on an individual candidate's total component mark of any bias, subjectivity, severity or leniency in an examiner's marking can be expected to be greater on paper, where a sole examiner is responsible for that total component mark, than on screen, where there are typically as many examiners as there are questions on the paper<sup>1</sup>.

Amongst other factors, segmentation contributed to feedback from the GCSE English examiners that was not altogether positive. They also spoke of becoming reluctant to stop to create an annotation box on screen to type in a comment, and so they used this facility

---

<sup>1</sup> It follows that segmented e-marking is less advantageous for a component made up of just one or two LFA questions.

sparingly. Together with the suggestion of loss of marking reliability, this feedback gave pause to ideas of drawing LFAs into marking on screen.

The analysis of EAR data from June 2005 and 2006 undertaken by Taylor (2007) represents another approach to exploring the impact of e-marking on quality of marking in GCSE written components. The most important finding from this analysis was that components marked on paper in 2005 and e-marked in 2006 had significantly fewer mark changes following a re-mark when they were marked electronically than when they were marked on paper. The analysis was repeated the following year for the components introduced to e-marking in 2007, including two components (in Religious Studies) that contained longer response items than had previously been e-marked). This time there was not a significant decrease but rather a slight increase in the number of mark changes following a re-mark for these components when they were e-marked, but this was non-significant, suggesting no negative effects of the introduction of e-marking (Taylor, 2008).

It was decided that, before planning any further marking quality and reliability research studies, it would be useful to carry out a qualitative exploration of how examiners respond to marking LFAs on screen. The exercise would aim to gain a greater understanding of the viability of marking LFAs on screen by listening to examiners' observations on the process as they marked, and looking for any suggestions as to why marking LFAs on screen might prove problematic, or any evidence that marking on screen using CMI+ could interfere with the usual cognitive processes involved in marking long written response. The following specific research questions were formulated.

- How do examiners experience marking responses of varying length on screen with CMI+?
- What, if any, limits are there to marking longer length responses on screen?
- What different or additional features would facilitate the marking of long written responses on screen?

## **METHOD**

### **Selected methodology**

Instead of depending on the usual interview and questionnaire approaches to gaining examiner feedback, it was suggested that the alternative, more direct approach of the 'think aloud' method should be used to address examiners' experience of marking responses of varying length on screen with CMI+. The method has very early origins in psychology as a means of investigating cognitive processes. It was dismissed by behaviourist psychologists because they were interested in empirical studies that could be replicated, and processes that could be observed from the outside, rather than events that take place internally in conscious thoughts, but the acquisition and analysis of verbal data or protocols using the method was promoted and further developed in the 1970s, most prominently by Ericsson and Simon (1980, 1993). The 1993 revised edition of their book, which was first published in 1984, points to a large increase in the use of verbal data to study cognitive processes. The method involves collecting and analysing verbal data from participants while they are carrying out a task so that the data are concurrent, drawing on short term memory, and therefore, unlike post-task interviews and questionnaires, not affected by changes that can result from retrospection. In the words of Ericsson and Simon (1993), collecting concurrent verbal reports "*rule(s) out the possibility that the information (the participants) retrieve at the time of the verbal report is different from the information they retrieved while actually performing the experimental task*" (p.xii). They hold

that direct reporting from short term memory without editing, explanation or theorising results in reliable data that limits the demands on brain processes.

Guidelines for implementing the think aloud method and analysing the resulting verbal protocols are given in detail in van Someren, Barnard and Sandberg (1994). They advise researchers to explain to the participants the nature of the task and then give them very general instructions about talking aloud as they work through it, using the same explanations and standard instructions to optimise reliability (Ericsson and Simon, 1993). Any further interactions with the participants should be neutral and unobtrusive in order to not influence their thoughts, the exception being to give a simple reminder to keep talking if they appear to be lapsing.

Ericsson and Simon (1993) recommend that the research issues being addressed are made explicit when using the think aloud method. Knowing the researcher's interest should help the participants feel more comfortable in carrying out the task, and more relaxed about being recorded. In the present context this means making the participating examiners aware that the researchers were interested in their experiences of marking longer answers on screen than in current AQA practice, and in particular on whether the volume of written response could become too great for them to feel confident in the accuracy of their marking.

Ericsson and Simon (1993) acknowledge that concurrent thinking and reporting may overload the participant and result in incomplete or lost information, with data possibly biased in the aspects that are attended to and reported. The investigation was planned in the hope that the examiners when thinking aloud would not be overloaded in touching on both aspects of marking on screen as well as features of the LFAs and the marks that should be awarded.

In the most recent relevant study using the think-aloud method, Suto and Greatorex (2008) report on its use in the context of script marking. Groups of six examiners from each of two GCSE papers, one of which used a points-based marking scheme and the other a levels of response marking scheme, talked aloud through their marking of a sample of scripts. Analysis of the verbal protocols obtained from the examiners supported the researchers in postulating a model of different and distinctive cognitive marking strategies. In a separate paper, Suto and Greatorex (2006) explored examiners' views on the think aloud method from their participation in the study of GCSE marking strategies. They found that the examiners had diverse views on whether or not the method had affected their marking, with some viewing it as a valid means of gathering information and others raising concerns about interference with their marking processes. The authors offer some explanations in terms of individual differences, and speculate that *"it might be that only certain individuals within a population will make suitable participants in a particular study"*, and suggest that *"task-piloting and participant-screening are necessary where the method is to be used in new contexts"* (p24).

Cotton and Gresty (2006) also reflect on the use of the think aloud method, in the context of a study of students' use of online learning resources. Their choice of the think aloud method was influenced by research into human-computer interactions and 'usability testing' of new software, which reinforces the idea that it should be suitable for finding out how examiners fare in marking LFAs using the CMI+ application.

The talk aloud method in the present context, of examiners seeking to satisfy themselves that they can mark LFAs on screen with a level of accuracy in line with that achieved for marking on paper, has a high degree of validity. The examiners were required to talk aloud as they marked on screen, about whatever they were thinking as they marked. Since they were negotiating a

variety of on-screen features offered by the CMI+ annotation tool it is likely that their thoughts would focus on how they were experiencing the software, and on the interaction between reading and assessing candidates' work and translating the assessment into the annotations and marks that the software would associate with the script image. All their comments, observations and verbalised thoughts were recorded, and these utterances were later transcribed for the purpose of a 'protocol analysis'. For such an analysis the utterances in each transcription are broken down into segments and the content of each segment classified according to an appropriate coding scheme.

### **The selected LFAs and their mark schemes**

A choice of LFAs in two different subjects was made with mark schemes that do not tie down how each mark might be gained, as in a points-based mark scheme, but instead give descriptions of levels of response to match against the responses. The questions and their associated mark schemes are included as Appendix 1. It will be seen that each of the mark schemes breaks down the total mark under a number of sub-headings. This is typical of the majority of LFAs, as has been confirmed in a subsequent audit of AQA LFA mark schemes.

The selected LFAs were:

- a GCE A level Biology essay in an A2 unit with a mark allocation of 25 marks awarded under four sub-headings (from Specification B A2 Paper BYB678/B);
- a GCSE French essay with a mark allocation of 20 marks awarded under three sub-headings (from the Specification A Higher tier Writing paper).

The Biology essay was allocated five pages at the end of a 16-page question paper booklet, following a section of short answer questions. The largest of the four sub-headings attracted 16 of the 25 marks. The French paper was made up of two LFAs each allocated 20 marks, with the selected essay also at the end of a question paper booklet, and allocated three pages. Both mark schemes have conventions with regard to annotations on the script. For example, the French essay set the candidate four 'tasks', or points to cover in the response, and the mark scheme required the examiner to annotate with 'T1' to 'T4' when s/he identified that one of these four tasks was being addressed. Both choices of LFAs gave the researchers the opportunity to select answers of varying length and numbers of words, because candidates could write as much (or as little) as they wished by including continuation pages if needed and tagging them to their question paper booklets.

### **The scripts**

About 100 scripts in each subject were selected and sent to DRS for scanning and presentation on screen. The Biology responses selected for e-marking ran from three to over six pages in length while the French responses ran from two to five pages. Some examiner annotations from the first marking of the scripts on paper were visible, but only faintly because they are written in red ink, which gives a much fainter screen image than the black ink which candidates are instructed to use. Fully 'clean' scripts would have required either photocopying with a red filter prior to scanning, which would have reduced the quality of the images, or greater attention to the scanning of each individual script, which would have proved too demanding on time and resources. The early scripts that featured in the familiarisation stage of the exercise were given this attention. Over-marking of the shadowy original marks was not seen as a particular issue since the exercise was not designed to investigate marking reliability *per se*. The examiners knew that, although they should make every effort to award accurate marks on screen, their marks were not to be used subsequently for any comparison purposes.

## **The participants**

A random selection of eligible examiners from the 2007 marking panels, all within range of AQA's Manchester office, was identified and the first three to accept the marking invitation for each of the selected papers were recruited. The Principal Examiner for each paper was also invited to participate and for Biology was able to accept. None of the participants had experience of using CMI+ with the annotation tool feature, although two of the Biology and one of the French examiners had marked electronically with CMI+ on a different component.

## **Procedure**

Seven examiners including the Principal Examiner for the Biology paper attended for the individual morning or afternoon marking sessions. They were made comfortable at the computer, in an empty room with a researcher sitting in on the session but not communicating during the recorded part other than to give an occasional reminder to talk aloud; no other prompts were given.

Each examiner was informed that the focus of the research was on how they were experiencing marking LFAs on screen, and in particular on whether the volume of written response could become too great for them to feel confident in the accuracy of their marking. They were made aware that AQA has been cautious in introducing LFAs to e-marking but was interested in identifying different or additional software features that would facilitate it, and that it was hoped that the examiners could suggest improvements. They were firstly given time to re-familiarise themselves with the marking scheme. Each was then introduced to CMI+ with the annotation facility and given time to become familiar with the tools available. Some LFAs were then presented for practice in navigating and marking on screen. When the examiner indicated readiness a further practice period was given for talking aloud while marking. Eventually, when they were ready, the recording began, and continued for about an hour, with CMI+ presenting different responses for each examiner to mark in the normal way.

For consistency, as required by the think aloud method, the same set of instructions was read to each examiner before they embarked on their marking. The instructions set the scene by acknowledging that AQA had to date stopped short of including LFAs in e-marking. The wording was as in Appendix 2 for the GCSE French component; minor changes were made for GCE Biology.

Further input from the examiners was sought about key aspects of marking on screen in an informal though partly structured post-marking interview covering the three research questions. The examiners were encouraged to suggest any improvements and any other ways to support the marking process in their subject in the light of all their marking. For example, they might have ideas for improving the readability of the responses on screen through such factors as the formatting and layout of the screen, different use of colour, location of the mark scheme, etc. To assist in generating such ideas the examiners marked a further selection of responses for an hour or so after the think aloud period was complete, with the annotation tool available for their use throughout. For Biology this additional marking could include responses to the short answer questions on the paper by way of contrast, but the French paper offered only the other essay question on the paper.

## Recordings

A video recording was also made of all the examiners' on-screen cursor movements using software that also recorded their spoken comments<sup>2</sup>. In this way it was hoped adequately to capture their experience of marking responses of varying length on screen with CMI+.

Each recording lasted at least one hour and the examiner spoke into a headset microphone. Unfortunately, the audio recordings for three examiners suffered some unexplained breaks, presumably through some malfunction of the headset microphone. What was recorded can be regarded for the analysis of the protocols as representative of the whole session however, since the recordings contained recurring talk-aloud content from one essay to the next. For one of the French examiners the audio recording most unfortunately terminated after five minutes and did not resume. This examiner could not be replaced because it was the beginning of the winter term and a substitute could not be found in the short window of time that DRS could make available for the investigation.

## RESULTS

### Examiner response to the talk aloud method

It is perhaps no surprise that the examiners, being drawn from the teaching profession, without exception quickly adjusted to talking aloud as they marked. The researcher in the room did not pay attention to what was being said, only to whether there were any breaks in the verbalisations. The prompt to 'think aloud' proved to be rarely needed.

The content of the verbal data given by each examiner differed, being consistently and predominantly of one of two kinds. The first, given by one of the Biology examiners and the Biology Principal Examiner, are essentially running commentaries on their actions on screen (not "*rambling discourses*" as Cotton and Gresty (2006) feared the method might deliver when they used it), with detail of their annotations and how they were adapting to e-marking. In the second kind, the examiners' verbal data show them to be fully focussed on what the candidate has written, on the content of each response, and what marks it should attract. The latter verbal protocols, as in the think-aloud research carried out by Suto and Greatorex (2008), would give an insight into their marking strategies but unfortunately are not very useful for analysis in respect of the aims of the current research.

The subsequent interviews suggest that the examiners in this latter group quickly adapted to on-screen marking and turned their full attention to the content of the LFAs (and possibly also the comments and marks of the original examiner). Their concern was to award, and to be seen to be awarding, the correct marks, leaving a complete audit trail - just as they do in their normal marking on paper – and this directed their spoken thoughts rather than reflections on their on-screen experience.

### Protocol analysis

The investigation gave very useful protocols for analysis from two of the Biology participants, who operated completely independently but expressed very similar thoughts as they marked. An analysis of their verbal protocols has been carried out to identify all the recorded 'thoughts' that are relevant to the use of the CMI+ software in marking LFAs. This is essentially a content

---

<sup>2</sup> TechSmith 'Camtasia' Studio software (version 4.0.2, 2007) was used for the video and audio on-screen recordings. Playback shows the computer screen with all the cursor movements and examiner inserts (marks, comments, abbreviations, underlinings, etc.).

analysis, as might be gained from analysing interview or questionnaire texts for example. It gives rise to the list below, which is included to show the breadth of the content categories identified in the recordings of these two examiners. The list is in an approximate order of frequency but there are no counts because it is obviously not possible to generalise from only two cases. Some explanatory notes are given in italics.

1. Pausing to create an annotation box (*right click on mouse*).
2. Having to leave an annotation box in order to scroll down to read more of the LFA (*i.e. beyond what's visible on screen*).
3. Having read more, returning to the annotation box to re-key its contents (*a comment could not be edited but instead had to be over-written, and this caused some concern since a candidate's response later in the essay might mean that the annotation had to be replaced*).
4. Scrolling back up, having read and annotated the whole LFA, to decide what mark to assign under each sub-heading (*i.e. pulling together all the annotations relevant to each sub-heading*).
5. Pausing to create an annotation box but reluctant to stop reading the LFA to do this.
6. Deciding to keep typing (and typing errors) to a minimum by curtailing comments.
7. Pausing to sideline irrelevant phrases.
8. Pausing to underline significant phrases or sentences.
9. Pausing to note a repeated error of communication.
10. Scrolling steadily down through the LFA - but wanting to navigate around the LFA faster.
11. Avoiding keying a number into an annotation box (*spelling it out instead: an instruction not to **start** an annotation with a number, because it will be picked up by CMI+ as a mark, was mistakenly taken as an instruction not to enter a number **anywhere** in an annotation box*), so that marks are entered only in the final reckoning under each sub-heading at the end of the LFA, as prescribed in the mark scheme.
12. Having difficulty locating the sub-heading mark and associated comment in the appropriate boxes at the end of the essay (*these boxes appear in the question paper booklet at the putative end of the LFA, and were scanned. The examiners all felt they should record their marks and comments in these boxes just as they would record them when marking on paper*).
13. Feeling some discomfort – stiff neck, fidgety.
14. Aware of slight blurring of image.
15. Needing a more sensitive 'zoom in and out' facility.

Interestingly, both examiners at some point in their marking recorded that they were '*losing the will to live*'. No doubt poor handwriting and a weak essay structure contributed here, together with the artificiality of the experimental marking and its location, but this comment was made in relation to the mechanics of using the CMI+ annotation feature. It seems that fairly small details can contribute to an element of irritation when e-marking.

### **Interview responses**

As is evident from the Introduction, the aims of the research allow for the possibility that the length of an LFA, whether in terms of the number of words to be read or the number of screens to be scrolled through, could be a factor in whether marking would transfer from paper acceptably. The post-marking interview with each examiner ranged over the three research questions:

1. Was marking the essay on screen manageable?
2. Do you think there are limits to how long the essay could be for marking on screen?
3. What different or additional features would make it easier to mark long written responses on screen?

As noted earlier, the French examiners had marked only LFAs while the Biology examiners had marked answers to short questions as well as LFAs in this investigation, but all the participants gave a positive response to the first question, and none indicated any limit to how long the essay could be for marking on screen. They were all clearly of the opinion that they had read and annotated the on-screen responses in the same way regardless of how long they were, and that the length of answer did not affect their ability to review their annotations and give sub-total marks over the whole answer according to the mark scheme. None of the examiners reported encountering LFAs of such length that they felt unable to mark them, indeed there was barely any comment on the length of individual essays. They had concerns as to legibility and structure, but not length *per se*. It seems that they take on the good, the bad and the middling just as they do on paper.

Although appreciating the automatic adding up of marks that is a benefit of e-marking (extended with the CMI+ annotation tool to wherever on screen the marks are entered), the examiners differed in their degree of enthusiasm for using the annotation tool, and in their opinion of how using annotation boxes etc might have interfered with and disturbed the marking process. It is clear that any developments by DRS to make it easier to use the CMI+ annotation facilities will be welcomed, and some are already in hand for the next release of the software.

### **Video recordings**

As noted earlier, the on-screen movements of the cursor were also all recorded, giving a full video record even where the audio recording was deficient. The recordings have therefore been passed to DRS as there is scope in the video records for further analysis for which they are best placed, for example of cursor movements to quantify the amount of scrolling. The recordings are time-consuming to watch and listen to, and even more time-consuming to analyse, but they will give valuable insights into how examiners use CMI+ and might usefully be undertaken again, especially when the next release from DRS of CMI+ with an upgraded version of the annotation tool is available.

## Overview of findings and observations

It was clear from the participants' recordings that examiners respond in individual ways to the task of marking on-screen. There was little commonality in the verbal records collected as the examiners marked, but the following points emerged in the recordings, from which some illustrative comments have been drawn, and from the interviews. They are grouped under the headings of the three research questions.

### 1. **How do examiners experience marking responses of varying length on screen with CMI+?**

The main finding was that the examiners seemed able to mark the LFAs on screen, regardless of how much the candidates had written. The Biology essay responses selected for scanning and marking ranged from over three pages to over six (including continuation sheets tagged to the answer booklet) while the selected French essay responses ranged from three to five pages. There was little by way of comment in the recordings on the length of the responses and on the evidence of the interview responses it appeared to be no more an issue than it is with marking on paper. The challenges of marking LFAs on screen exist regardless of how much a candidate has written. Deciphering difficult hand-writing is one of these, as it is on paper, although there was a comment that the appearance of a response is not exactly the same after scanning:

“for all I don't doubt it's a brilliant scanner, it is not as clear, quite, as writing appears on the page. It's got just a slight blur to it, very slight but it's there...”

### 2. **What, if any, limits are there are to marking longer length responses on screen?**

The research did not find any limits to response length. The examiners adapted the paper-based mark scheme as necessary, in order to annotate the response on reading through, for example using side-lining for irrelevancies (Biology) and symbols to represent ticks, and then proceeded to work through all the scripts presented on screen. A possible explanation of how the LFA marking task was undertaken successfully, for any length response, is as follows.

Typically a LFA mark scheme will break down the mark allocation into chunks, with the marks under different sub-headings collected together in the total. The examiners did as they do on paper, which is to revisit the comments and other annotations they have made on reading through the response, in order to decide on the sub-heading marks. They entered these marks at the end of the response to be added up by CMI+ (giving the correct total because no other marks were entered directly, although they may have been embedded in comments). The suggestion is made here that this breakdown of marks under a number of sub-headings is what makes marking LFAs feasible, because it breaks the task down into smaller manageable tasks. A consequence is that the CMI+ facility to gather up marks scattered around the response is not useful in marking LFAs that do not use a points-based mark scheme, because it cannot identify which sub-heading the marks belong to, and it does not allow the examiners to make considered judgements under each sub-heading. The sub-headings need to carry a manageable number of marks and it will be an empirical matter to establish how large each can be for e-marking a particular component. For example, the essay in GCSE English A (which was the topic of the earlier investigation of the annotation tool in CMI+ (Fowles, 2006b)) has sub-headings of communication and organisation, with 18 marks, and sentence structures, punctuation and spelling, with 9 marks. The examiners appeared to find it more difficult when

e-marking to apply the levels of response marking scheme and award the marks reliably; the marking was shown to be considerably more reliable in a paper-based exercise.

### **3. What different or additional features would facilitate the marking of long written responses on screen?**

Several of the suggestions for different or additional software features, with illustrative comments, have been collated in Appendix 3. All suggestions have been passed on to DRS to categorise, in consultation with AQA staff, as '*must haves*', '*nice to haves*' or '*not feasible*', with appropriate priority ratings, where they are not already being actively developed for DRS's next release of the software.

## **DISCUSSION**

It was disappointing that the examiners' verbal data were not all complete due to unexpected and unexplained technical problems. However what was recorded contained a degree of repetition that means it was probably representative of the full session. A greater problem was that the recordings were not rich, other than for two of the participants, in the hoped-for output of thoughts pertinent to navigating the LFAs on screen, and in shedding light on the specific questions posed for the research. This point has already been mentioned and explained as probably reflecting on examiners' ease of transition to marking on screen, but some alternative methodological approaches will be discussed later.

There is no reason to suggest that either the Biology or the French LFAs in the exercise are untypical of LFAs. Although they were both allocated space in a question paper booklet for the candidates' responses, which is not typical of LFAs (they are more usually given free-form space in a blank answer booklet), they appeared at the end of the booklet and could run onto additional tagged pages. There is therefore no reason to suggest that candidates' responses would have differed in a free-form answer booklet, for example that they might have written at greater length. The mark schemes were also typical of LFAs in that they specified a unique, subject-specific set of abbreviations/annotations for examiners to use in identifying particular features of the LFA. Further, they both break down the total mark under sub-headings which again is typical of many LFA mark schemes, as an audit of current and prospective (in new specifications) mark schemes has subsequently established. This means that the task is broken down into smaller parts carrying proportions of the total marks which could make the on-screen marking task more manageable. This suggestion regarding LFA marking on screen is discussed further later.

Although the categories from the protocol analysis of the verbal data (listed earlier) might be seen as having a negative slant and therefore reflect poorly on e-marking LFAs, this would be to misread the e-marking experience. The examiners were directed to thinking about their marking in the new context of electronic marking and so they were obviously more concerned about differences and potential problems rather than noting how readily they could adapt to e-marking, or other positive aspects. For example, each examiner marked the LFA presented on screen using the normal practice of identifying the features to reward in the answer as they read, reviewing the answer at the end and deciding what the marks should be under each of the prescribed sub-headings. Seeing the computer's total of these marks, delivered in a box at the bottom of the screen without the examiner having to do the arithmetic, got an enthusiastic response when mentioned in the post-marking interviews, even though that enthusiasm was nowhere registered in the recordings and would not therefore appear in the protocol analysis.

## Methodology and alternatives

The questions in this research exercise addressed particular aspects of LFA marking, as outlined in the Introduction, and it was hoped that the verbal data from the talk aloud method, from examiners saying whatever was in their minds as they marked, would include some of relevance to these questions. It is not to deprecate the examiners or the talk aloud method that this did not always prove to be the case in their verbal protocols. While it was a positive finding that the examiners were able to maintain their verbalisation while e-marking, the talk aloud method did not permit the researchers to intervene to steer the content of the verbalisation. By the same token, the prompt when needed was only to 'talk aloud', not for example to ask for an explanation of why verbalisation had ceased.

Other researchers have reported that the method can fail to elicit the verbal data they expected or give data with reduced informational value, *"possibly fuelling misconceptions about their usefulness and trustworthiness"* (Leighton, 2000, p11). Leighton warns against collecting verbal reports from tasks that (a) *"evoke automatic processes"* (and are not amenable to verbal description) or (b) *"overload working memory and stymie verbalisation"* (p13). As already noted, Ericsson and Simon (1993) acknowledge that verbal data may well be biased in the aspects that are attended to, and also that concurrent thinking and reporting may overload the participant and result in incomplete or lost information. In this exercise the primacy of the marking task appeared to stymie verbalisation of the 'nuts and bolts' of how it was proceeding.

The obvious implication is that the think aloud method may need to be supported by different techniques. For example, Suto and Greator (2008) used the retrospective method of a post-marking interview. Taylor and Dionne (2000) also complemented concurrent verbal protocols with a retrospective approach, in an investigation of problem solving that compared professors and students. They were satisfied that using both approaches gave data that *"accessed a broader spectrum of problem-solving strategy knowledge and demonstrated more patterns of difference across expertise (professors and students) than did either method used alone"* (p419).

However Kuusela and Paul (2000) carried out an experiment, involving decision making and choice behaviour, in which they collected both concurrent, think-aloud data and retrospective questionnaire data. They report that the *'concurrent data provided more insights into the decision-making steps'* (p387), and that they obtained more verbal data for analysis concurrently than retrospectively. They interpret this finding as confirming that *'short term memory fails over time and that subjects are more capable of verbalizing their task-related thoughts if protocols are collected during the choice task'* (p397). They comment that another factor that can be evident in retrospective data is that participants may be sensitive to the researchers' expectations and *'may report their actions and thought processes in a socially more desirable fashion'* (p391). By the same token, it might be suggested that participants could have a personal agenda connected with the task in hand and try to manage the impression they give to the researcher as much in their concurrent as in their retrospective verbalisations.

In another approach, Cotton and Gresty (2006) decided to have a range of prompts to help in collecting the type of data from participants that they were looking for because they had reservations about whether the simple instruction to 'think aloud' would give *"meaningful data or merely a rambling discourse"* (p48). Their range of prompts was limited however, and designed only to guide participants into useful verbalisations. They describe their elaboration of the

method as "a kind of 'prompted think aloud' with the aim of encouraging students to articulate their thinking as clearly as possible and to enhance the data collected" (p 50).

Methodological variations are also suggested by van Someren *et al* (1994). They compare the talk aloud method with a number of alternative methods: those involving *observation of behaviours*, resulting in 'action protocols' for analysis; *structured techniques* which direct the responses required of participants and includes questionnaires, with either closed or open questions; and a third category of *verbal reports*, which yield verbal protocols. In this latter category they identify five methods: *introspection*, where the participants are asked to give their own accounts of a cognitive process as accurately, fully and coherently as they can; *retrospection*, where they are questioned retrospectively about their actions and thoughts; *using questions and prompts*, where the participants are interrupted by the experimenter for explanations of what they are doing or thinking; and *dialogue observation*, involving an exchange of information about actions and thoughts.

The comparisons of these methods by van Someren *et al* are made with particular reference to whether they (a) disturb performance of the process or task which is the subject of the investigation, (b) suffer from memory loss, where there is a time delay between carrying out the task and giving responses on it, and (c) are subject to reinterpretation by the participant. They advocate the think aloud method because normally it is unlikely to interfere with the task, it involves only short term memory, and it does not give the participants time to reinterpret their performance. They go on to suggest that it might nevertheless be appropriate to address a particular research issue using two or more alternative methods in combination, particularly an alternative method in conjunction with the think aloud method. As noted earlier, Suto and Greatorex (2008) and Taylor and Dionne (2000) both combined the think aloud method with *retrospection*, while Cotton and Gresty (2006) combined the think aloud method with *questions and prompts*.

An alternative strategy, that would capitalise on the teaching background and inclination of examiners, would be to combine the think aloud method with what might be labelled an *observed dialogue* session, by asking an examiner to explain to a novice examiner how to go about the marking task in the light of his /her own experience on screen, with the novice asking questions in what would be a natural learning and dialogue situation. The dialogue could be observed by the researcher and s/he could also analyse the verbal data in an audio recording. The teaching examiner's marking is of course interrupted, as it is with *questions and prompts*, but probably more fruitfully as the novice needs to imitate the teacher and will therefore encourage appropriate verbalisations.

### **Suggestion regarding LFA marking on screen**

The 'think aloud' exercise and the post-marking interviews supported the suggestion, mentioned earlier but yet to be validated, about what makes marking LFAs on screen viable. This suggestion concerns the mark scheme structure, and it is hoped that it can assist AQA in drawing up a programme for drawing components with increasingly complex LFAs into e-marking.

The on-screen marking of each of the selected Biology and French essays was well supported by its mark scheme, which in both cases breaks down the total mark into a number of sub-headings. As the examiners worked through a response they would use the prescribed annotations as appropriate, and then scroll back at the end to review all their annotations and to allocate the sub-heading marks according to the rules and/or descriptions of levels of response in the mark scheme. It is suggested that in this way the marking task is broken down and made

manageable, with a coded audit trail in place via the annotations. This is in contrast to a mark scheme that gives overall, multi-faceted descriptions of different levels of response, requiring the examiner firstly to judge which description best matches the essay, and secondly to assign a mark from the range available to the selected response level. There is a much higher memory load for this holistic judgement than where the mark scheme requires the examiner to identify, and attend to, the different facets of the LFA separately. This suggests that the demands of marking LFAs on screen can perhaps be placed on a demand continuum determined by the number of sub-headings and the marks available for each. The suitability of an LFA for e-marking may be considered with this in mind: an LFA that is broken down into parts, none of which attracts more than 10 marks, would appear to be at the 'less demanding' end of the continuum, and thus more readily considered for e-marking, while an LFA at the other end of the continuum, allocated a large number of marks distributed amongst only a few levels of response, would appear to be at the 'more demanding' end. An example of a question paper with LFAs at the latter end of the continuum can be seen in a current AS English Literature unit, where questions are allocated 35 marks distributed between six levels of response. Another example, from GCE Media Studies (Unit 6), is a question allocated 50 marks with a 'levels of response' mark scheme that identifies 10 marks for each of 5 levels. This may be more demanding and less reliable to mark than where the 50 marks are allocated to several different sub-headings and levels of response identified within each.

One of several new CMI+ developments by DRS will allow different facets of an LFA to be side-lined in different colours. Sections of the response sidelined with the same colour can be moved by the software to be adjacent to each other. The response and annotations can then be reviewed in a logical order for marking without having to scroll around the complete LFA. This is at yet untested but would appear to be a very useful aid to support examiners' marking, by using different colours to identify aspects of an answer, for instance those relevant to the different mark scheme sub-headings. This is a development for LFAs that should enhance marking reliability - but it is only available as a feature of electronic, as opposed to paper-based, marking.

### **Suggestions for further research**

- (a) Lengthy responses were selected for this investigation, and very generous scanning 'clip areas' were defined to ensure that the longest were all captured. To progress further with electronic LFA marking it will be necessary to identify the LFA's complete location. This might mean introducing defined response areas for candidates to use for their answers, which is likely to have an effect on the amount written: candidates are influenced by how much space is allocated for their answer, and treat it as a clue to what is expected, just as they treat the number of marks allocated to each question printed on question papers as a pointer to how much 'worth' an answer will be given. The two papers in this study are presented to candidates as a question paper booklet with provision for tagging extra pages (which for e-marking cause complications as they cannot be scanned normally). The impact of introducing a defined response area on the length and quality of answers produced by candidates should be recognised, for example in the script appraisal of the awarding meeting, as a significant change to what is required of candidates. The alternative to defined response areas is some form of electronic recognition of the beginning and end of a LFA, and also its identity (i.e. question number), which would leave candidates still free to determine how much to write as is currently usually the case, i.e. in a free-form answer booklet rather than a question paper booklet.
- (b) AQA's planned programme of transferring increasingly complex LFAs from paper-based to electronic marking, alongside DRS's development of the e-marking software, will be

accompanied by regular research input in carrying out marking reliability studies. These will be post-marking exercises of the kind carried out for GCSE English (Fowles, 2006b), providing a comparison between e-marking and the equivalent paper-based marking. As these are expensive and time consuming studies to carry out, the selection of the (few) components to include will be important because they will be used to generalise to similar components with LFAs marked on-screen. Should a quantitative marking reliability study demonstrate any difficulties in maintaining the quality of marking, the qualitative and exploratory approach of the think aloud method might again be used. Alternative concurrent approaches might be used to explore examiners' actions, for example the *prompted talk-aloud* of Cotton and Gresty (2006), *question and prompts* (van Someren *et al* (1994)), or the *observed dialogue* method suggested above, in which an examiner is asked to explain and introduce marking on screen to a novice examiner, with the novice asking questions in what would be a natural learning and dialogue situation. This method would mimic the talk aloud method in terms of the analysis but in the present context would allow the researcher to intervene to steer the verbalisation onto the 'nuts and bolts' of e-marking and away from detailed justification of the marks awarded. A concurrent method is probably best complemented by a retrospective approach, and it is suggested that the *observed dialogue* method could be accompanied by a separate structured interview with each participant (examiner and novice).

## CONCLUSION

The exercise has suggested a continuum of demand in marking LFAs which it is hoped will have some validity as DRS develops the e-marking software further in subsequent releases of CMI+ and its annotation tool, and AQA rolls out a planned programme of transferring increasingly complex LFAs from paper-based to electronic marking. A regular research series of quantitative marking reliability studies is planned to accompany this programme of change. Should a quantitative study demonstrate any difficulties in maintaining the quality of marking, the qualitative approach of the think aloud method might again be employed, perhaps combined with other concurrent and retrospective approach(es). As a concurrent method, the think aloud method offers direct and undisturbed feedback on examiners' performance which is not subject to any memory loss or any reinterpretation as a result of any time delay between carrying out the task and giving responses on it. A variation of the think aloud method might be the suggested approach of an *observed dialogue* with a novice examiner.

Dee Fowles

May 2008

W:\Dee\e-marking LFAs\ LFA report v4

## REFERENCES

- Cotton, D. and Gresty, K. (2006) Reflecting on the think-aloud method for evaluating e-learning. *British Journal of Educational Technology*, v37 n1 p45-54
- Ericsson, K.A., and Simon, H.A. (1980) Verbal Reports as Data. *Psychological Review*, v87 n3 p215-251
- Ericsson, K.A., and Simon, H.A. (1993) *Protocol Analysis. Verbal Reports as Data (revised edition.)* Cambridge, Massachusetts: MIT Press

- Fowles, D. (2002). *Evaluation of an e-marking pilot in GCE Chemistry: Effects on marking and examiners' views*. AQA Research Committee paper, RC190.
- Fowles, D. (2005). *Evaluation for QCA of the CMI+ system in the January 2005 e-marking live pilot*. AQA Research Committee paper, RPA\_05\_DEF\_WP\_10.
- Fowles, D. (2006a) *How reliable is the marking in GCSE English?* AQA Research Committee paper, RPA\_06\_DF\_RP\_020
- Fowles, D. (2006b) *How well does marking in GCSE English transfer to marking using CMI+ with annotation?* AQA Research Committee paper, RPA\_06\_DF\_RP\_047
- Kuusela, H. and Paul, P. (2000) A Comparison of Concurrent and Retrospective Verbal Protocol Analysis. *American Journal of Psychology*, n113 v3 p387-404
- Leighton, J.P. (2004) Avoiding Misconception, Misuse, and Missed Opportunities: The Collection of Verbal Reports in Educational Achievement Testing. *Educational Measurement Issues and Practice*, Winter, p6-15
- Raikes, N. (2002). *On screen marking of scanned paper scripts*. Cambridge: University of Cambridge Local Examinations Syndicate (UCLES).
- Suto, I. and Greatorex, J. (2006) *What do GCSE examiners think of 'talking aloud'? Interesting findings from a preliminary study*. Paper presented at BERA Annual Conference
- Suto, W.M.I. and Greatorex, J. (2008) What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process *British Educational Research Journal*, v34 n2 p213-233
- Taylor, K.L. and Dionne, J-P. (2000) Accessing Problem-Solving Strategy Knowledge: The Complementary Use of Concurrent Verbal Protocols and Retrospective Debriefing. *Journal of Educational Psychology*, v92 n3 p413-425
- Taylor, R. (2007) *The impact of e-marking on Enquiries after Results*. AQA Research Committee paper, RPA\_07\_RT\_TR\_050
- Taylor, R. (2008) *The impact of e-marking on Enquiries after Results, June 2008*. AQA Research Committee paper, RPA\_08\_RT\_TR\_019
- van Someren, M.W., Barnard, Y.F. and Sandberg, J.A.C. (1994) *The think-aloud method. A practical guide to modelling cognitive processes*. London: Academic Press. Retrieved 4 March 2008 from <http://staff.science.uva.nl/~maarten/Think-aloud-method.pdf>.

## SELECTED LONG FORM ANSWERS AND MARK SCHEMES

## GCE Biology Specification B 2007 A2 Paper BYB678/B

**S 4** Write an essay on **one** of the topics below.

**EITHER**

(a) Movements inside cells.

**OR**

(b) Transfers through ecosystems.

*In the answer to this question you should bring together relevant principles and concepts from different parts of the specification.*

*Your essay will be marked not only for its scientific accuracy, but also for the selection of relevant material.*

*The essay should be written in continuous prose.*

*The maximum number of marks that can be awarded is:*

<i>Scientific content</i>	<i>16</i>
<i>Breadth of knowledge</i>	<i>3</i>
<i>Relevance</i>	<i>3</i>
<i>Quality of Written Communication</i>	<i>3</i>

**MARK SCHEME****General Principles for marking the Essay:**

Four skill areas will be marked: scientific content, breadth of knowledge, relevance and quality of language. The following descriptors will form a basis for marking.

**Scientific Content** (maximum 16 marks)

Category	Mark	Descriptor
<b>Good</b>	16	Most of the material reflects a comprehensive understanding of the principles involved and a knowledge of factual detail fully in keeping with a programme of A-level study. Some material, however, may be a little superficial. Material is accurate and free from fundamental errors but there may be minor errors which detract from the overall accuracy.
	14	
	12	
<b>Average</b>	10	Some of the content is of an appropriate depth, reflecting the depth of treatment expected from a programme of A-level study. Generally accurate with few, if any, fundamental errors. Shows a sound understanding of the key principles involved.
	8	
	6	
<b>Poor</b>	4	Material presented is largely superficial and fails to reflect the depth of treatment expected from a programme of A-level study. If greater depth of knowledge is demonstrated, then there are many fundamental errors.
	2	
	0	

**Breadth of Knowledge** (maximum 3 marks)

Mark	Descriptor
3	A balanced account making reference to most areas that might realistically be covered on an A-level course of study.
2	A number of aspects covered but a lack of balance. Some topics essential to an understanding at this level not covered.
1	Unbalanced account with all or almost all material based on a single aspect.
0	Material entirely irrelevant or too limited in quantity to judge.

**Relevance** (maximum 3 marks)

Mark	Descriptor
3	All material presented is clearly relevant to the title. Allowance should be made for judicious use of introductory material.
2	Material generally selected in support of title but some of the main content of the essay is of only marginal relevance.
1	Some attempt made to relate material to the title but considerable amounts largely irrelevant.
0	Material entirely irrelevant or too limited in quantity to judge.

**Quality of language** (maximum 3 marks)

<b>Mark</b>	<b>Descriptor</b>
3	Material is logically presented in clear, scientific English. Technical terminology has been used effectively and accurately throughout.
2	Account is logical and generally presented in clear, scientific English. Technical terminology has been used effectively and is usually accurate.
1	The essay is generally poorly constructed and often fails to use an appropriate scientific style and terminology to express ideas.
0	Material entirely irrelevant or too limited in quantity to judge.

**Total 25**


## GCSE French Specification A Higher tier Writing paper

**Q2** You have seen this article about the school of the future in a magazine.

**Le Collège de l'Avenir**

Pour aider les jeunes dans une société changeante à réussir dans l'avenir, est-ce que les collèges doivent changer? Si oui, qu'est-ce qu'ils doivent considérer?

- **les facilités?**
- **les matières?**
- **la vie sociale?**
- **les visites scolaires?**



Qu'en pensez-vous? Ecrivez-nous avec vos opinions.

Ecrivez une réponse en **français** et donnez vos opinions et les raisons pour vos opinions.

Mentionnez:

- les facilités nécessaires pour un collège idéal
- les matières qui sont essentielles pour l'avenir **et pourquoi**
- l'importance des amis au collège
- une visite scolaire que vous avez faite.

(20 marks)

### MARK SCHEME

This question will consist of *four* tasks and will be assessed for Degree of Communication, Range/Complexity and Accuracy, giving a maximum of 20.

## DEGREE OF COMMUNICATION

- The assessment of Communication depends upon recognising an attempt to write something about the task. All responses should be self-contained and comprehensible without reference to the rubrics.
- Candidates will be expected to go beyond the minimum level of response in order to score at least 2 marks. This means that the candidate will have added an additional piece of information, in the form of a phrase or clause (hereafter called a **Development**), which goes beyond the minimum required by the task.
- For a Development to be accepted, the quality of language should be the same as that which is required for a Communication task.
- Developments cannot be credited for a task which has been rejected.
- There can be *up to two* developments per task in this question.

### Tasks to Marks - Degree of Communication

Tasks	Marks	Requirement	Degree of Communication
0	0		Nothing of merit; fails to communicate OR occasional words are recognisable within sentences but no complete messages are communicated.
1 – 4 (0 Dev)	1		Communicates <i>a little</i> basic information (e.g. simple facts).
1 - 4 (1 Dev)	2		
2 - 4 (2 Dev)	3		Some basic information is conveyed; occasional additional details conveyed (e.g. description, simple opinion).
2 - 4 (3 Dev)	4		
3 - 4 (4 Dev)	5	Must include an opinion, if not, revert to 4 marks	Communicates clearly <i>quite a lot</i> of relevant information, including personal opinions; regularly goes beyond a basic response to give more detailed information relating to descriptions and accounts.
3 - 4 (5 Dev)	6	Must include an opinion, if not, revert to 4 marks	
4 (6 Dev)	7	Must include justification of an opinion, if not, revert to 6 marks	Communicates <i>a lot</i> of relevant information; candidate can narrate events, give full descriptions and can express and justify ideas and points of view.
4 (7 - 8 Dev)	8	Must include justification of an opinion, if not, revert to 6 marks	

## QUALITY OF LANGUAGE

- Marks will be awarded out of 6 for each of Range/Complexity and Accuracy. The marks will be added to make a total out of 12 for Quality of Language.
- The mark awarded under Range/Complexity must not be more than **one mark** higher than the mark awarded for Degree of Communication.
- The mark awarded under Accuracy must not be more than **one mark** higher than the mark awarded for Degree of Communication.
- If a mark is awarded for Communication this will inevitably lead to the award of a mark for Range/Complexity and for Accuracy.
- For the award of 4 marks or more under Range/Complexity there must be at least one reference to two of past/present/future events.

<b>Range / Complexity</b>	<b>Marks</b>	<b>Accuracy</b>
Very little effective vocabulary. There are occasional recognisable words but they make little coherent sense.	<b>0</b>	There is little, if any, evidence of understanding of the most basic linguistic structures.
The vocabulary and structures used are simple, often repetitive, limited in range and may contain many cognates.	<b>1</b>	There is only limited understanding of the most basic linguistic structures and most sentences contain major errors.
Vocabulary is appropriate to the basic needs of the task. Structures are simple, often repetitive and are rarely linked.	<b>2</b>	Most sentences contain errors, many of a major nature, and verb forms are rarely accurate.
Vocabulary and structures are appropriate to the task with a little attempt at variety and there is some successful attempt to link structures together.	<b>3</b>	There are some major errors and frequent minor ones. Attempts at verb forms and tense formations are often unsuccessful.
There is some variety in the use of vocabulary and some successful attempts at a variety of structures including attempts at longer sentences using appropriate linking words. Some personal opinions are successfully expressed. There are successful attempts at using more than one time frame.	<b>4</b>	There are a number of minor errors and a few major ones, but the piece is more accurate than inaccurate. Verb forms and tense formations are not always correct, but the intended meaning is clearly recognisable.
There is a wider range of vocabulary and structure which communicates descriptions and opinions with some precision. Longer sentences, including the use of subordinate clauses, are used more regularly and with increasing success.	<b>5</b>	Inaccuracies are mainly of a minor nature although some major errors may occur when complex structures are attempted. Verb forms and tense formations are usually correct.
A wide range of vocabulary and structures appropriate to the topic is effectively used. Longer, more complex sentences are handled with confidence producing a fluent piece of coherent language.	<b>6</b>	There are hardly any major and few minor errors even in more complex structures. The overall impression is of accuracy and verb forms and tense formations are secure.

## INSTRUCTIONS READ TO THE GCSE FRENCH EXAMINERS

We have chosen this French paper and invited you here today to help us, in collaboration with DRS, to move forward with e-marking.

The two questions on this paper perhaps make it appear unsuitable for e-marking and, with 20 marks available for each, the answers are certainly much longer than anything that is currently being marked on screen. We want to know how manageable you find them, especially the really long ones. We want to explore with you today what would help make e-marking more suitable for your essays, primarily in terms of the way the screen is set up and the actions you have to take to carry out the marking.

The way we want to do this is by asking you to **think aloud** as you mark the answers presented on screen. Thinking aloud means asking you to say out loud everything that you would say to yourself silently as you mark. Just carry on as if you were alone here and speaking to yourself.

The pc will record everything you say, so we would ask you to speak clearly. It will also record what's happening on screen and of course this will help us afterwards to make the link with what you've said.

As you'd expect the answers vary in their length and readability and this will probably influence how you think about them and, crucially for us, what you say as you mark them.

If you are silent for any period of time we'll remind you to keep talking by just saying 'keep talking' – we appreciate that it won't feel too natural, at least at first, and that you may get caught up in the French and forget that what we're actually interested in here is as much, if not more, to do with the mechanics of the marking as in the marks themselves.

We obviously aren't going to use your marks for real, nor could we use them to make a valid comparison of e-marking with paper-based marking because there are only a few scripts and a few examiners involved, but we would ask you to mark as normally and as accurately as possible.

Do you understand what you're being asked to do here? Do you have any questions?

When you're ready we'll move on and ask you to talk aloud as you mark and we record.

## Suggestions for different or additional CMI+ features to facilitate the marking of long written responses on screen

### 1. Marks under sub-headings, and in prescribed positions

It has been suggested that splitting the marks allocated to marking an LFA between sub-headings is likely to make the task more manageable. However it does raise the design issue of whether these marks should be recorded in pre-determined locations, and if so whether these should be on the screen layout or printed on the scripts, as in the Biology paper, where S to Q are the mark sub-headings:

END OF QUESTIONS		
For Examiner's use only		
	Mark	Comment
S		
B		
R		
Q		

The examiners put marks in the boxes in this grid, as they do on paper, although they found it hard to centre them in the box, and they appeared somewhat wobbly. The French examiners entered three sub-heading marks horizontally at the end of the response, in line with their normal practice on paper, and these also appeared somewhat wobbly as they found it difficult using the mouse to position them in a straight line. The Biology grid above is 'For examiners' use only', but on paper it acts for candidates as a pointer to expected response length. This might be better avoided, and in e-marking the grid could be replaced by boxes on screen, visible to the marker throughout their marking. One examiner commented:

"it'd be handy actually if you could, if the mark box was a...sort of the box you clicked into and then put your marks in, rather than just putting the pointer where you want to put the marks...if you could select one of the mark boxes like S, B, R, Q, click it then just put (in) the mark..."

This suggests that a screen addition might be the facility to limit the mark collection function of the annotation tool to marks entered within a pre-determined number of boxes, perhaps near the total mark box, with the mark instructions designating as many boxes as required for the marks under the specific sub-headings of the mark scheme. The maximum mark under each sub-heading could be set, as it is for the total mark. The examiner quoted above also suggested having a box 'floating' on screen, and always visible, displaying the comments and annotations.

Another reason to prefer specified boxes for collecting in the marks, if the mark scheme lends itself to this, is that double-clicking to insert a single mark can happen by mistake. If the

examiner notices this s/he will delete the mark, otherwise it stays in the total and will be an error that is not picked up. Further, starting an annotation with a number puts that number into the total, which might also not be picked up as an error.

## 2. Scrolling

Scrolling back through the response after having read and annotated it was found to be clumsy and difficult.

'Here I've got to scroll through...there's no page skipping...there's the usual bar for scrolling through on the right hand margin but there isn't the page skip which I'm used to, to get to the end'.

In fact marking probably requires a number of 'sweeps' through the response, particularly if there are different sub-headings to attend to, and easy navigation is essential e.g. the facility to move instantly back to the top or to the bottom, or to select pages via thumbnails. Thumbnails would also give an immediate indication of the length of the response, which is possibly useful. DRS are actively addressing these navigation issues.

The annotations are in a different colour but need to be large enough to stand out when scrolling through. Examiners found it best to use the right hand margin for annotations they knew they would need to refer back to.

## 3. Annotations

- (a) The main irritant in using the annotation tool was having to stop to create a box, through a right click on the mouse, and then keying in either an accurately typed comment or an abbreviation. If the comment contained an error it would need over-typing to correct it. It is less demanding to enter an abbreviation than a wordy comment and fortunately several abbreviations are prescribed in the mark scheme, for example 'Q' to indicate a lapse in the quality of writing in the Biology mark scheme (Appendix 1). The French examiners are also given a set of abbreviations to use (for example, 'T1' to 'T4' as mentioned earlier) and it is fairly typical for mark schemes to instruct examiners to leave letters, numbers and symbols by way of shorthand to indicate what they have identified in a response (but note that the ubiquitous tick is not currently available through CMI+).

- (b) Annotations are a little awkward to insert.

"I think making these boxes is going to be a pain really. It would be so much better if you could just put a 'D' (straight) in - obviously it would be quicker wouldn't it than having to do all these clicks, right click and left click..."

and the process could be inhibiting:

"a reasonable A-level content' - that's what I would write at the moment (as a comment but not being a great typist I could just see me ending up saying 'reasonable', which wouldn't be very informative going back to centres to be honest".

There was some comment that having to stop and make a box before annotating might interrupt the marking process. On paper examiners are less conscious of stopping to write a comment,

whereas online it causes a break in their thoughts. They have to physically make the annotation box and cannot simply start typing.

"normally we put in a letter Q to show things like spelling mistakes and so on ...and there's no way I'm actually going to stop each time, create a text box, put in Q, forget what I was reading and go again... so I think things like Q and underlining wrong things are just out cos it takes too long".

- (c) Examiners found it frustrating that they have to make an annotation box in order to enter a mark greater than 9. For marks of 9 or less they quickly become practised at keying the number directly onto the screen.
- (d) Inadvertently starting a comment with a number in an annotation box puts that number into the total, which could introduce an error.
- (e) Annotations are in a small font size and, although in colour, they do not stand out and could be missed in reckoning up the sub-heading and overall marks.
- (f) A particular concern was to have the facility to update an annotation. Situations where this would be desirable include (i) correcting a long comment that contains typing errors and (ii) updating a mark embedded in an annotation after reading further on down the response. Examiners were frustrated by not being able to revisit their comments and marks.

"Ah...I meant to put an asterisk there so I can't amend that, so I've got to delete it...put a new box in".

- (g) A related issue is that some sort of drag and drop function to re-locate the annotation box would be useful – if a comment is in the wrong place it has to be deleted and re-typed elsewhere. A 'copy and paste' facility might also be useful for repeated comments.

"A drag and drop feature so I could make a box and pull it exactly where I want it to be...that would also encourage me to annotate exactly by the piece of relevant information which I think would be more helpful".

- (h) Also related is a need for a 'wrap round' facility for annotations. The current annotation tool restricts an individual comment to a single line. As part of the audit trail and to justify the mark awarded, the examiner may find a single line inadequate.