National Foundation
for Educational Research

An Evaluation of the 2001 New Technologies Pilot

A report for the Qualifications and Curriculum Authority

Paul Newton, Chris Whetton, Ewan Adams, Jenny Bradshaw
and Carolyn Wong

FULL REPORT

7 November 2001

Acknowledgements

The authors are very grateful to the many individuals who contributed to the project:

NFER Project Team

Dougal Hutchison, Tilaye Yeshanew, Margaret Parfitt

Additional NFER staff

Joan Howell, Simon Rutt, Pauline Benefield

QCA Project Team

Graham Hudson, Helen Patrick, Andy Cleasby, Barry Creasy, James Butler

NCS Pearson Project Team

Martyn Leese, Kris Knowles and colleagues.

The Marking Teams

Jayne Boaler, Margaret Cooke, Chris Driver, Brian Speed and all members of the English and maths marking teams.

Additional Stakeholder Participants

Tim Cornford, Neil Pirie, Steve Addison, Malcom Britton

Contents

Contents (cont.)

Contents (cont.)

Contents (cont.)

Contents (cont.)

Section 1 Introduction

Since its inception in the mid-1990s, the external marking of national curriculum tests has been challenged on numerous fronts. Critics have argued that it is costly, error-prone, a burden for teachers and that it fails to deliver results rapidly. The External Marking Agencies have also raised concerns that it is becoming harder to recruit a sufficient number of appropriately qualified markers.

The QCA has acknowledged these problems and continues to take steps to address them. Toward this end, the most significant development of recent years has come about as a result of the Invest to Save Budget submission (Number 62) from QCA, ACCAC, CCEA and SQA to HM Treasury.

1.1 Invest to Save

The Invest to Save Budget (ISB) submission was proposed with the intention of improving the speed and efficiency, accuracy and value for money of the UK's national testing systems – ultimately, the hope was that this would enhance public confidence in national assessments throughout England, Wales, Northern Ireland and Scotland. Through a series of eight related projects the ISB submission proposed ways in which improvements might be delivered in relation to the four main stages of the testing process: design, administration, marking and results data collection.

These eight projects were specified as follows:

1. on-line marking (centralised or web-based);
2. on-line training of markers;
3. personalised scripts for schools;
4. change management;
5. management information and reporting;
6. interface with DfEE's Information Management Strategy Project;
7. develop management framework;
8. on-line delivery of test papers.

1.2 The 2001 New Technologies Pilot

The intention of the 2001 New Technologies Pilot (NTP) was to begin taking forward the projects of the ISB through a small-scale exploratory study conducted in the relatively safe environment of a non-statutory national test. The environment chosen was the Year 7 progress tests.

1.2.1 The Year 7 Progress Tests

The Year 7 progress tests (Y7PTs), in mathematics and English, were formally introduced in 2001 with the intention of monitoring the progress of children who had failed to achieve level 4 in those subjects in the previous year. Although administered to pupils toward the end of Year 7, the tests were based upon key stage 2 (KS2) programmes. In fact, for 2001, the Y7PTs were exactly the same as the KS2 tests (with the exception of their covers).

Although the tests were not statutory in 2001, all schools with eligible pupils were encouraged to register for them. In fact, not all schools did, and this was partly a consequence of other parallel initiatives which involved the administration of alternative tests for Year 7 pupils. Yet a large number of eligible pupils were still entered for the Y7PTs during 2001. (A *Headteacher's Declaration Form* return report, from 18 June 2001, recorded the participation of 2311 schools for English, 2358 schools for maths and 3240 schools overall.)

A sub-set of schools that entered pupils for the Y7PTs was selected to take part in the NTP.[1] Although participating in the Pilot had very little impact upon these schools, there were a number of subtle procedural differences. First, they received scripts directly from the NTP contractor rather than from the EMA. Second, the scripts received were personalised with the names of pupils who were to sit the tests. Third, they received scripts back from markers only after they had also been processed by the NTP contractor.

---

[1] The method of sample selection for the Pilot was somewhat convoluted. From the 2365 schools that had registered an interest in January 2001, a significant number were removed (overseas schools, schools in the KS3 strategy, schools in the Y7PT pilot during 2000, schools with entries for only one subject, and schools with timetable variations). From the 1655 remaining schools, 50 were selected at random from each of the three regions. From these 150, a further random sample of 111 schools was selected (apparently to reach a rough total of around 5,000 participating pupils).

This processing involved guillotining the scripts (for scanning) and recombining them (after scanning).

1.2.2 The NTP specification

The Pilot was intended as an exploratory study of the degree to which novel approaches to marking and data collection – facilitated by developments in the fields of information and computing technology – could successfully be applied in the context of national curriculum testing. Such approaches have been employed successfully in other countries, notably the USA, and the challenge was to determine whether they could be equally effective in the UK.

The main characteristic of the novel approach under investigation was the manner in which images of pupils' scripts were electronically scanned into a central computer and then distributed to a team of 'e-markers' who would mark the script images on-screen using software delivered over the internet. There are many potential benefits of such an approach, for example:

1. marking might be speeded up, as scripts would spend less time passing between school, EMA and marker via the postal system;

2. marking might become more accurate, as markers could focus upon specific questions rather than having to mark all questions from a paper;

3. as a result, questions that did not require subject matter expertise to mark could be marked by non-experts, thereby using human resources far more effectively;

4. the data collection would become entirely automated, eliminating the need for mark sheet completion and mark sheet data entry and all of the inevitable errors that creep into these processes.

The remit of the 2001 NTP was not to implement all of the projects within the ISB, but to focus on those most directly concerned with the technical innovations described above. Thus, the Pilot was to revolve predominantly around projects 1 to 3 and 5, that is: on-line marking and training technology, personalised scripts and management information.

The Project Initiation Document (PID) for the NTP specified a number of objectives for the prospective contractor. These were: "...

1. to ensure that the project is managed to time, quality and cost as set out in the specification and contracts;

2. to develop and pilot a system for scanning, imaging and marking on PCs the scripts of a sample of pupils for the year 7 progress tests in English and mathematics which is scaleable to a national level;

3. to agree on test paper designs which NCS can produce for this pilot and that would be suitable for scanning (without changing the questions or test layout as set already);

4. to agree on guidance to be used for future test paper designs that would produce paper[s] to use for high performance scanning;

5. to identify schools (with QCA) which will participate in the new technologies pilot exercise;

6. to appoint, train and support a sufficient number of suitably qualified persons as 'e-markers';

7. to provide new technologies pilot schools and 'e-markers' with a telephone support service which deals with their queries as described [in] the relevant service level agreement;

8. to mark test script images for new technologies pilot schools making use of the mark scheme agreed nationally and ensuring that prescribed measures for accuracy are met;

9. to scan and process test scripts sent to the pilot agency in a timely fashion so that their return to the Y7EMA marker is not delayed (turnaround times to be agreed);

10. to produce accurate test results for each school and pupil in the new technologies pilot, together with pupils' results profiled at the item level. (The results will not be sent to schools as part of the new technologies pilot, only used as part of the evaluation.);

11. to provide the NDCA with data to meet the requirements as set out in the DfEE information requirements for key stages 2 and 3 as they apply to Y7PT;

12. to identify interfaces with all other agencies involved in the year 7 progress tests and develop service level agreements to govern the performance of these interfaces;

13. to provide suitably and appropriately detailed information (through full and open participation) to the new technologies pilot evaluation project to allow for effective assessment of the future deployment of technology and the marking and data collection process."

In fact, certain requirements of the PID were amended subsequent to this specification. For example, it was agreed that a remote system of web-based marking via the internet would be replaced by a local system of centre-based marking via an intranet. This meant that markers would not need to be provided with a telephone support service. It was also agreed that data would not ultimately be provided to the NDCA, but provided instead to the QCA and to the evaluators.

In order to safe-guard the marking of the Y7PTs, it was decided that all Pilot scripts would be marked through the conventional process before being marked using the new technology process. This also had the benefit of providing data against which the new technology marking could be compared. The conventional marks were captured by the NTP contractor and analysed by the NTP evaluator. Importantly, the conventional marks were not available to the e-markers during the new technology marking process.

1.2.3 The NTP contractor

QCA appointed NCS Pearson as the 2001 NTP contractor.

NCS was formed in 1962 in Minneapolis, USA. It is a global information services company, providing software to collect data, assess, score, mark and provide statistics from this data, mainly in the education market.

In September 2000, NCS merged with Pearsons, an international media company based in the UK. The new company was named NCS Pearson.

1.3 The 2001 New Technologies Pilot Evaluation

To provide an independent analysis of the extent to which the Pilot was successful, the QCA developed a tender specification for an evaluation contract. The basic aim of this evaluation was to determine the strengths and weaknesses of the Pilot processes, in terms of their impact upon the speed, accuracy and reliability of data capture and transfer.

1.3.1 The evaluation specification

The evaluation PID outlined the following broad objectives: "...

1. to evaluate whether or not the pilot operated as specified;

2. to evaluate the speed, accuracy and reliability of pupils' results produced by the pilot processes when compared with conventional marking;

3. to evaluate markers' attitudes to using the new technologies;

4. to provide measures which can be used to validate the original business case for investigating this approach to marking;

5. to provide an evaluation of the possible ways in which results data could be provided to, and used by, schools, pupils and parents with a recommendation for a preferred format;

6. to provide a considered opinion on how well the processes might scale-up to implementation for all national curriculum tests;

7. to identify any major constraints which might act as barriers to implementation of the processes for all national curriculum tests;

8. to provide recommendations for the improvements of the process identifying key areas which must be addressed before a scaled-up system could be implemented."

Again, once the evaluation contract had been awarded, certain requirements of the PID were amended. The main two changes were the omission of the fifth objective and the inclusion of an additional objective concerning a technical problem that was identified when the Pilot went live: "...

9. to investigate the frequency with which pupils' responses to questions ran beyond the 'clip image areas' ".[2]

---

[2] Pupils' responses were stored in two separate formats. First, full-page images of pupils' scripts were stored. This was predominantly to enable electronic versions of pupils' scripts to be sent back to schools. Second, discrete images of pupils' responses to individual questions were stored. This meant that, for each question on each paper, 'clip image areas' were defined within which it was assumed that pupils' responses would lie. These item response images were what markers saw when they marked pupils' work. Problems occurred when pupils had responded beyond the clip image areas, as these out-of-image responses were not accessible to markers.

It is important to note that the objectives of the trial implied analysis at a number of distinct levels. These can be specified as four over-arching goals:

1. to **evaluate** whether the NTP contractor managed successfully to implement the agreed procedures for 2001;

2. to **evaluate** whether the procedures implemented during 2001 were effective in delivering significant benefits without undue costs;

3. to **consider** whether the procedures implemented during 2001 might be scaled-up for all national curriculum tests to deliver significant benefits without undue costs;

4. to **consider** whether revised procedures for future years might deliver significant benefits without undue costs.

The principal focus of the evaluation was taken to be goals 1 and 2, bearing in mind that the ultimate purpose of the Pilot was to guide future practice, as suggested by the more reflective goals 3 and 4. The main point for the evaluators to take into account was that certain of the specific procedures implemented during the 2001 trial would not necessarily be employed in future years if implemented on a larger scale.

1.3.2 The evaluation contract

To achieve the objectives set out in the revised specification, the NFER proposed a series of seven inter-linked studies. The studies involved both formal and informal interaction with managers, staff and markers at NCS Pearson as well as interaction with other key stakeholders. They involved direct and indirect access to the procedures being implemented and analysis of qualitative and quantitative information and data arising from the project. The seven studies are described below.

1.3.2.1 Study 1: Preliminary interviews with stakeholders

Study 1 was intended to provide the background and scoping information required for the study as a whole. Interviews were conducted with the main stakeholders in the process to gain their perspectives on the Pilot, its operation and what it could achieve. These interviews were underpinned by the collection and study of relevant documents and procedural protocols. Interviews were conducted with:

1. a DfES Senior User on the Joint Project Board;

2. a QCA Senior User on the Joint Project Board;

3. the NCS Pearson Senior Supplier on the Joint Project Board;

4. the Project Manager for NCS Pearson;

5. the National Data Collection Agency Senior Supplier;

6. the External Marking Agency Senior Supplier.

The interviews were audio-recorded, and notes of these were returned to the participants for correction or changes of emphasis.

1.3.2.2 Study 2: Questionnaires and interviews with markers

Study 2 focused on the attitudes and experiences of the markers. All markers were asked to complete two questionnaires, one during the early stages of the marking and the other after marking had been completed. The initial questionnaire gathered information on the experiences of the markers, the training received, their aspirations for the Pilot and their proposed strategy for undertaking their marking. It also posed questions such as whether the pay structure seemed fair and attractive.

At the end of the process, the second questionnaire examined markers' experiences of on-line marking, the difficulties encountered, how they had structured their work, views on home working, and also revisited issues such as pay.

The questionnaires were supplemented with a series of individual interviews, held over the course of the marking period. All markers were interviewed at least once. The interviews elaborated on issues covered in the questionnaires, probing them more deeply, especially in relation to the nature and organisation of the work and the operation of supervision.

1.3.2.3 Study 3: Observation of the scanning and marking process

Over the course of the marking period, an observer from the NFER was present for long periods in the marking centre. There were two elements to the observations, structured and unstructured. For the structured observation, a checklist was developed of the activities of the markers and of activities and events within the centre more generally. At set times during the day, a record was made of the currently occurring activities. This supported an evaluation of the nature of task demands, the efficiency of processes, the

levels of supervision and management required, and markers' attitudes (as evinced by their application to the work in hand).

The unstructured observation was based on the observer being present at the marking centre and recording salient events as they occurred. The purpose of this element was to obtain an impression of the efficiency of the overall process and to be present as unforeseen events unfolded, allowing the methods of recovery to be noted. Such unstructured observation also supported a deep understanding of the processes of the Pilot and the extent to which its aspirations were achieved.

1.3.2.4 Study 4: Collection and analysis of management data

One of the aims of the evaluation was to determine the kinds of management data that would arise through the Pilot, to collate them, and to compare them with comparable data from conventional marking. This involved working closely with managers at NCS Pearson and QCA to determine the kinds of data that could be accessed and to specify the format in which it would be most useful.

Management data were collated within a number categories, which included:

- speed of scanning (i.e., elapsed time and scanning rates during active periods);
- marking rates (i.e., elapsed time for all scripts; number of person hours expended; working rates for different marker types);
- quality control information (i.e., number of 'seed' items presented; accuracy rates per marker; supervision interventions; error logs).

All measures were collated separately for mathematics and English.

1.3.2.5 Study 5: Collection and analysis of measurement data

NCS Pearson were able to provide item-level data not only from the new technology marking, but also from the conventional marking of the same scripts. These data were compared using a variety of statistical techniques including correlation, mean mark difference analyses and concordance analyses.

NCS Pearson also provided data on the marker monitoring functions supported by their software. These included:

- between-marker reliability (AKA 'reliability');
- within-marker reliability (AKA 'consistency');
- absolute-marker reliability (AKA 'validity').

These data were analysed to determine a baseline of marking reliability that might be expected from the new marking technologies, to explore differences between papers and questions, and to consider differences between markers.

1.3.2.6 Study 6: Reflective interviews with stakeholders

A few months after the main elements of the process were completed, a second set of interviews were undertaken with the same participants that were involved in Study 1. Within these interviews, participants were invited to reconsider their aspirations for the Pilot in the light of feedback that they had received. In particular, views were sought on the advisability of proceeding with an introduction of on-line marking for all national curriculum tests, and on the way in which implementation of change might best be achieved.

1.3.2.7 Study 7: Frequency of responses beyond the clip image area

With the assistance of software developers at NCS Pearson, it was possible for an NFER researcher to view full-page images of pupils' scripts remotely through a browser. Overlaid on these images were the clip image areas. Using a sample of 300 pupils, and viewing their scripts for all six test papers, an analysis of the frequency with which pupils' responses ran beyond the clip image areas was conducted.

1.4 The evaluation report

The following sections present the results of the seven studies outlined above. Their findings overlapped to some extent, but they have been presented independently for the sake of thoroughness.

Section 2 explains the procedures that were agreed for the 2001 Pilot and the way in which they were (or were not) implemented. This section is based predominantly upon the observations of Study 3 and from procedural documents relating to the Pilot.

Section 3 considers the views of markers that were involved with the process during 2001. It presents data arising from Study 2.

Section 4 is concerned with the views of stakeholders and, as such, presents findings from Studies 1 and 6.

Section 5 presents the quantitative data related to management and measurement concerns. It is, therefore, concerned with the results of Studies 4 and 5.

Section 6 outlines the extent to which pupils recorded responses beyond the clip image areas that NCS Pearson had specified and, as such, is based upon the results of Study 7.

Finally, Section 7 presents a synthesis of the findings that emerged from the seven inter-linked studies, with particular reference to the four goals identified earlier.

Section 2 The implementation of agreed procedures

The aim of Section 2 is primarily to achieve the first of the four over-arching goals specified in Section 1:

1. to evaluate whether the NTP contractor managed successfully to implement the agreed procedures for 2001.

This will be achieved using evidence gathered from across the seven studies, but primarily focusing upon findings from Study 3 (the observation). The following presentation begins by explaining the procedures that were employed during the Pilot, from beginning to end, and continues by considering those aspects that were most and least successful.

While the evaluation team did its best to penetrate the subtleties of operation it is possible that some areas were not examined in sufficient depth. As such, there may have been particular strengths or weaknesses that were not fully appreciated and that are not fully acknowledged in the following pages. Indeed, certain of the more technical issues concerning the operation of the software and the scanning of the scripts are likely to have evaded full scrutiny. However, it is hoped (and felt) that the major issues of relevance to the Evaluation were uncovered.[1]

2.1 The functioning of the Pilot

Section 2.1 describes the major procedures of the Pilot as they actually functioned. Where actual procedures contrasted with intended procedures this is typically mentioned but is considered in greater depth in Section 2.2.

2.1.1 Answer booklet modification

To enable pupils' scripts to be scanned effectively, NCS Pearson had to make certain modifications to the standard Y7PT answer booklet scripts. These included:

1. adding scanner 'timing' marks, and page identifiers, to the margins of each page of each script;

2. adding personalised barcodes to each script, indicating school and pupil IDs;

---

[1] The following discussion has benefited significantly from comments provided by NCS Pearson in relation to an advance copy of this section.

3. printing school and pupil names on the front of each script.

The personalisation of scripts represented a significant departure from conventional practice. Traditionally, scripts are not personalised until pupils write their names on the front of their answer booklets at the beginning of each test. Advance personalisation does have the disadvantage of an additional stage in the process and requires significant co-ordination between the new technology contractor and other agencies responsible for collating school and pupil information. In addition, requires teachers to ensure that pupils receive the correct papers prior to testing. Yet, in addition to enabling effective scanning, it has other benefits such as reducing a burden upon teachers who are no longer required to sort completed scripts in alphabetical order. Personalisation also ensures the most effective tracking of confidential scripts.

2.1.2 Software preparation

The software for the trial was based upon the generic NCS Pearson Distributed Workflow System (DWS), run through the NCS Pearson UK intranet. DWS was the foundation software of the new technology process, which managed the component processing sub-systems. DWS was a pre-existing NCS Pearson system, but was customised to meet the needs of the Pilot. It was anticipated that this software would also form the basis of future marking contracts won by NCS Pearson.

While the basis of the system was developed in-house, the component used to perform the expert marker function – Netgrade – was leased from a third party (in China). While the Netgrade function appeared quite compatible with DWS, NCS Pearson explained that it would not be used in future marking contracts. Instead, the expert marker function would utilise NCS Pearson's in-house system – ePEN. Apparently, the costs of employing ePEN for the 2001 trial were prohibitive and this function would only be cost-effective on a much larger scale.

2.1.3 Despatch of test papers and answer booklets to schools

After the personalisation of answer booklets had been completed, they were despatched to schools. This process was managed through the DWS Mail Manager program. Within each school batch, answer booklets were individually logged out. Thus, the precise despatch content, time of despatch and despatch administrator were logged for each script of each batch.

2.1.4 Test administration and conventional marking

The tests were administered by schools in the usual manner and, as usual, teachers sent the completed scripts to the conventional marker assigned to their school. After approval of their first sample, these markers made the Pilot scripts their priority. Once they had completed the marking of the Pilot scripts they sent them directly to NCS Pearson.

NCS Pearson commented that markers had not always sent scripts to them as quickly as had been anticipated (with knock-on effects for the scanning schedule). It was further suggested that this was because markers had experienced delays in receiving scripts from schools.

2.1.5 Script return to NCS Pearson

The script return process was managed through the DWS Batch Builder program. Individual scripts were logged in and collated into 'batches' (a batch was a functional grouping of scripts from a single paper, containing no more than 50). Once again, the precise receipt content, time of receipt and receipt administrator were automatically registered for each script of each batch.

The reconciliation of despatch with return was managed automatically through the DWS Report Server. Where schools had failed to enclose scripts that had been logged out, discrepancies were followed-up manually by telephone. This was sometimes difficult when schools had already broken up for the vacation. Batches with outstanding queries were put to one side and not sent directly to be guillotined.

For each batch processed, a paper form entitled *DWS Batch Header* was produced and was fastened to it.

2.1.6 Script scanning

Script scanning involved a number of discrete stages, from the initial guillotining, to the main scanning, to procedures for dealing with scripts that could not be scanned automatically.

2.1.6.1 Guillotining

Before scripts could be scanned, they had to have their spines (including staples) removed. This enabled scripts to be fed through the scanners as single pages.

One of two machines was used to guillotine each script: a 'slicer', that simply shaved approximately 0.5 cm from the margin of each booklet; and a 'power guillotine', that was more precise and removed less of the spine, but that required more time to operate. The latter was generally used only for the writing scripts, as they often had responses in the margins, and thus needed a minimum of the spine to be removed.

The slicers were located in the main (largest) hall, outside the hall in which expert e-markers worked. A number of data entry and clerical markers also worked in this area.

2.1.6.2 Scan Master

The process of scanning was managed primarily through the DWS Scan Master program. Once again, this logged the precise scan content, time of scan and scan administrator for each script of each batch. The barcode from the *DWS Batch Header* form was scanned in prior to the commencement of scanning for each batch and each page from each script within each batch was automatically reconciled.

NCS Pearson had two 5000i scanning machines, devoted to the scanning of scripts from English and maths respectively. NCS Pearson reported that these machines could scan at a rate of 9,000 sheets per hour at peak performance, although they normally planned on a rate of 5,760 sheets per hour. Neither rate was actually achieved in practice (the reasons are explained more fully in Section 5). Generally, each machine required only a single operator.

During the scanning process, two sets of images were captured. The first set corresponded to 'clip image areas' – the areas of each page in which pupils' answers were located. The second set corresponded to 'full-page images' – the complete image of each page from each script. The clip image areas are all that NCS Pearson would generally capture and the requirement for a double capture procedure was the principal reason why the scanners did not achieve the intended rate of 5,760 sheets per hour.

If a script did not go through the scanner on the initial attempt, it was re-fed. If it was again rejected, it became an 'exception' and was scanned by hand on a flat-bed scanner (see below).[2]

---

[2] A report is available on-line, incidentally produced by the DWS Report Server application, that shows the barcodes that are in the 'exception' process until such a time that the 'exception' has been resolved.

Reasons for a script being rejected by the scanner included the following.

- Pages torn, or in some other way damaged.
- Pupils (or markers) having written in the margin of an answer booklet such that the 'timing' marks were interfered with. This happened particularly in the case of writing responses, as pupils often wrote beyond the line and into the margin.
- A continuation sheet having been attached to a writing answer booklet.
- The use of answer booklets that had been photocopied by schools. These could not be automatically scanned as their ID numbers corresponded to different pupils.
- Answer booklets that had duplicated pages due to errors in the printing process.

2.1.6.3 Scan Master Exceptions

Scripts that could not be scanned automatically had to be scanned manually, using a separate flat-bed scanner. This process was managed through the DWS Scan Master Exceptions program (again, logging precise scan content, time of scan and scan administrator for each script of each batch).

Hand-scanning on the flat bed scanner required each sheet of each 'rogue' answer booklet to be scanned. This required very precise placing of the sheet to line up the timing marks. The Exceptions process captured only clip image areas. This meant that full-page images had to be captured separately. This was carried out using the DWS Scan Master Attachments program, again on a flat-bed scanner. The Attachments program was also employed for writing scripts where additional sheets had been attached (as additional sheets posed an obstacle for the automatic processing of scripts via the high speed scanners).

Scanning via Exceptions and Attachments was a very laborious and slow process. The member of staff responsible for Quality Assurance of batching and scanning estimated that a single 'rogue' answer booklet might take fifteen minutes to go through the hand scanning process. Once again, the flat-bed scanner generally required only a single operator.

2.1.7 Script despatch to markers

Once the full and clip images from all scripts within a batch had been successfully captured, the batch could be sent back to its original marker. However, before this could

happen each script had to be reconstructed. This was achieved with a staple through the top left-hand corner.

The script return to markers was managed through the DWS Mail Manager program (which logged despatch content, time of despatch and despatch administrator for each batch). At the time of despatch the DWS Report Server provided a record of the automatic reconciliation of script despatch against script receipt.

NCS Pearson had a four-business-day turn-around target for scripts; that is, it was intended that all batches should be scanned and despatched to markers within four days of receipt.

2.1.8 Item image administration

Corresponding to the two sets of images captured for each script, two separate sets of images of pupils' responses were stored. The first set contained full-page images for each page of each script. This was, essentially, a customised function for NCS Pearson and it was to serve three functions:

1. to enable data entry staff to record marks that had been awarded during the conventional marking process (note that, when scanning the full-page images, markers' annotations were recorded as well);

2. to enable supervising markers to access a full record of each pupil's work (in order to resolve queries that might arise when, for example, a pupil's response had run beyond the clip image area);

3. to enable electronic versions of each script to be prepared for return to schools (a process managed through the ISTORM program).

The second set contained only the contents of pre-defined areas within each script, corresponding to the spaces in which pupils were required to record their answers to each question. These item image clips formed the basis of the electronic marking model. The logic of the entire process was that answers to individual questions would be separated and stored individually. Then they would be distributed to markers individually and markers would mark responses to specific questions rather than marking entire scripts. The 'clips' were stored for distribution, and then distributed to markers, via the DWS Pre Edit program.

As the scanners were not optimised to read red ink, the 'conventional' Pilot markers were provided with a customised red pen which could be read effectively. However, as conventional markers recorded marks in the margins of each page – beyond the clip image areas – these were generally not visible to the 'on-line' Pilot markers (although occasional problems were encountered when conventional markers' wrote in other places – see 3.4.2).

2.1.9 Electronic marking

As noted above, the logic of the new technology was to mark at item level rather than at script level. The principal benefit of this is that markers can specialise. First, those with the most expertise can be assigned to only those questions requiring that expertise (while less skilled markers mark the less complex questions); second, markers can focus upon a smaller number of questions, for example, a single marker might be assigned to mark only responses to a single question.

2.1.9.1 E-marker structure

Markers were divided into three functional categories: response selection, clerical and expert. In an exactly parallel way, every question from each of the six Y7PT papers in maths and English was assigned to one of these three functional categories. Consequently, response selection markers marked only response selection questions, clerical markers marked only clerical questions, and expert markers marked only expert questions.

Note that response selection markers also performed the data entry function, inputting marks from the conventional marking process. Data entry, response selection and clerical marks were all captured through the DWS Editor program. Expert marks were captured through the Netgrade system.

2.1.9.1.1 Response selection markers

The simplest questions to mark were those with a limited range of possible responses. Response selection markers directly inputted pupils' answers and these were scored

automatically by a program that had been set-up to reward only those responses deemed correct by the Chief Marker.[3]

To be classified into this category, the possible answers for a particular question needed to be relatively simple, to ensure that they could be both input and scored unproblematically. Markers were required to refrain from inputting an answer whenever they were unable to interpret it clearly. In this situation, they would be required to 'sticky note' the clip, i.e., to send it directly to a supervisor, without inputting it, but with an electronic note explaining what their query was.

As the task of response selection was simply to input the responses present on each script (which were subsequently scored automatically) it was possible to appoint relatively unskilled staff to these positions. Responses to all response selection questions were input twice to ensure the integrity of the process.

2.1.9.1.2 Clerical markers

When a question stimulated responses that tended to be less well circumscribed, and therefore required a small element of judgement to mark, it was designated as clerical. Clerical questions tended to require judgement as to the degree of approximation of the actual answer to the intended answer, but tended not to require the professional judgement of an experienced teacher. As such, it was possible to appoint intelligent but relatively unskilled staff to these positions. Responses to all clerical questions were marked twice to ensure the integrity of the process.

When clerical markers were unsure what mark to award to a pupil's response, they were required to 'sticky note' it, i.e., to send it directly to a supervisor without marking it. The supervisor would then mark it. As previously, clerical markers were required also to explain, in their sticky note, exactly what their concern was. In their User Guide, they were told to restrict this to one of three messages: unable to read answer; unable to mark response; unable to see full image.

Clerical markers could have been required to mark on a single-question basis. If so, each marker would have started marking a particular question (say Q5 from the first maths

---

[3] The term 'response selection marker' was coined by the Evaluation team to distinguish this function from the input of conventional marks (NCS Pearson used the generic term 'data entry staff'). In fact, 'response selection' is something of a misnomer, as markers do not actually select responses from a menu.

paper) and continued marking responses to this question until there were no more responses to mark. S/he would then progress on to another item. This would not necessarily mean that *only* s/he would mark responses to that question; for example, all markers could be assigned to Q5 from the first maths paper until all responses had been marked, then they might all progress to the next question.

However, instead, clerical markers were required to mark on an question-set basis. This meant that they marked all responses to a set of clerical questions from pupil 1, then all responses to the same set of clerical questions from pupil 2, then all responses to the same set of clerical questions from pupil 3, etc.. Essentially, then, they marked in a manner similar to conventional marking – marking by pupil rather than by question – the difference being that they marked only clerical questions and, during each session, they would typically mark only a sub-set of all clerical questions. (Response selection markers, too, were assigned question-sets rather than individual questions.)

2.1.9.1.3 Expert markers

When a question stimulated responses that required the professional judgement of an experienced teacher to mark, it was designated as expert. As such, only experienced teachers were appointed to these positions. Responses to expert questions were marked only once, but there were a number of monitoring procedures in place to ensure the integrity of the marking process (see below).

When expert markers were unsure what mark to award to a pupil's response, they were instructed to seek assistance from a senior marker. If immediate assistance could not be provided then they were to provide a mark and, in addition, send a message to their supervisor (using the Netgrade messaging function) explaining their concern.[4] Messages concerning specific items would be sent to a supervisor for consideration. It was not entirely clear whether supervisors were required to give feedback to markers on the subject of each message, or whether they were supposed simply to decide whether the mark awarded to the item needed amending. Supervisors spent a lot of time in face-to-face discussion with markers and automated messages were not responded to

[4] Unlike the sticky note facility, the Netgrade messaging system was not directly linked to specific items. Instead it was a more general system for enabling communication between markers. If an expert marker had a query regarding a specific clip, the Encrypt ID and Item No. of the clip would have to be copied into the message manually and sent to a specific supervisor.

immediately. For this reason, feedback to markers on the concerns expressed in their messages would have been delayed by a significant period of time and, as such, not entirely optimal.

While response selection and clerical markers marked on a question-set basis, expert markers marked on a single-question basis. That is, they were allocated to a particular question and marked responses to that question until they were allocated to a different question. In fact, the allocation process was slightly more subtle than this, in that expert markers could be allocated to a single expert question for each of the papers that they were entitled to mark. Expert markers still undertook their marking on a single-question basis. Indeed, they were required to log on to a specific question before they began marking and, if they wished to mark responses from a different question, they were required to log off from the first and then log on to the second.[5]

The allocation process was undertaken manually by supervisors and administrators. Although this was done according to a default plan, markers were able to request to be allocated to a new question if they so desired. Unfortunately, the item allocation process proved more awkward than was intended, owing to the fact that markers ran out of responses to mark more rapidly than had been anticipated. This may have been attributable to a failure to scan scripts in sufficient time, or due to the speed with which markers were marking, or due to both. The amount of time required for re-allocating expert markers had the unfortunate consequence of distracting supervisors from their principal task of supervision. (In fact, marker allocation was not technically within the remit of the senior markers; they only assisted in this function due to the unexpected demand for re-allocation in the early stages.)

It is worth mentioning that expert markers were able to return to an item that they had marked previously using the Back Score function within Netgrade. If they so chose, they were able also to amend the marks that they had awarded. While the marker was logged on to a particular question, the Back Score function allowed her/him to return to any of the previous ten items marked.

---

[5] In fact, it became apparent that markers did not need to formally log off between items as double clicking the title bar when logged in would allow the marker to change to a different test paper or question item. However, the documentation and User Guide was not updated to reflect this.

2.1.9.1.4 Supervising markers

For each subject, two supervising markers were appointed: one Chief Marker and one Senior Marker. Although they took part in the routine e-marking to a small extent, their primary role was one of supervision, adjudication and allocation of expert markers (although the latter was not formally within their remit).

2.1.9.2 E-marker appointment

Appointments to the data entry, response selection and clerical marker posts were made by NCS Pearson, largely through employing temporary staff from the local area. Appointments to the expert marker positions were intended to be based on the recommendation of the External Marking Agency. However, as the EMA did not provide lists of markers very quickly, this was not always possible and markers were recruited by other means (e.g., word of mouth and local advertisements). While it was originally anticipated that the markers would be recruited from a national sample of experienced markers (who would reside in Rotherham for the duration of the marking period) the actual sample tended to be more local and tended to have significantly less external marking experience.

There were usually four data entry staff working in each of two shifts: from 6am until 2pm, and from 2pm until 10pm. In times of need, extra staff were recruited from other NCS Pearson divisions or from the other shift rosters. For example, for a few hours on the afternoon of 6 June, there were eight staff working to clear a backlog on data entry: one on marks and totals, three on (unspecified) data entry, and four on names.

The clerical markers tended to be young students, although one ex-maths teacher was also appointed to this function. Clerical markers also worked in two shifts: from 6am until 2pm, and from 2pm until 10pm. Average numbers on each shift were four or five working on maths, and one on English. Two NCS Pearson staff worked as adjudicators for clerical marking, but did not supervise (all supervision of clerical marking fell to the senior markers).

Seventeen English expert markers were appointed, of whom three did not attend due to illness or for other reasons. Nine maths expert markers were appointed. These were in addition to the two senior markers appointed in each subject.

2.1.9.3 E-marker training

Two main forms of training were undertaken:

1. training for expert markers (and supervisors) in the conventional marking of the Y7PTs for 2001;

2. training for all markers in the use of software at NCS Pearson.

Training in the conventional marking process was undertaken externally, as part of the national training of Y7PT markers, and organised by the Y7PT EMA. It took place on 12 May (English) and 19 May (maths), a few weeks before e-marking began at NCS Pearson. Training in the software at NCS Pearson was undertaken internally and was conducted separately for the different categories of marker.

Internal training for expert markers extended over a single day for each subject, on 21 and 22 May. Experience with computers varied greatly between members of the marker groups, with some English markers being particularly unfamiliar with the technology. Some had little familiarity with even the most basic of processes (e.g., logging in, using a mouse, working within windows or with a web browser). Following the morning training session, there was an opportunity to use the Netgrade software during the afternoon. This was not very structured and, despite a plenary session, there was little opportunity for detailed feedback.

Internal training for non-expert markers took place on the day that they started marking. It was not observed by the Evaluation team, but few problems were commented upon. This may have been due to the relative simplicity of the tasks involved and the youth of the temporary staff (with the implication of a greater familiarity with computers).

All markers were provided with software User Guides (and expert markers had also been provided with conventional marking handbooks).

2.1.9.4 E-marker working environment

All markers were based at NCS Pearson UK headquarters in Rotherham. The main tasks took place within two large, adjacent halls. The first hall housed the scanning machines and the expert markers. The second hall housed the slicing machines and the remainder of the markers. Each marker had a designated workspace and PC linked to the NCS

Pearson UK intranet. The workspaces were neither cluttered nor cramped. Markers were able to take formal or informal breaks when so desired.

Working conditions were generally very good, with spacious, well-lit areas, drinking water available, and some snack food facilities. Although there was no canteen, an area for breaks was provided. The centre had good parking facilities, but was a reasonable distance from any rail station.

2.1.9.5 E-marker administration

In addition to the supervising markers, administrators from NCS Pearson were on hand to ensure that the marking process ran smoothly. The administrative staff kept a watch on all aspects of the Pilot and took prompt measures to solve problems as they arose and to prevent their re-occurrence as far as possible. The general atmosphere in the workplace was friendly and positive, in part due to this responsiveness from the administrative team.

2.1.9.6 E-marking software

The DWS software appeared to function well, was attractive and generally user-friendly. The integration of the DWS and Netgrade systems also appeared to function effectively. Where problems were encountered the network development team were generally able to rectify them promptly and without serious loss of time or of data. (Although see 2.2.9.6 for a number of problems that were more recalcitrant.)

In addition to the standard marking screens, markers were also able to access on-line mark schemes. Indeed, they were advised to use these mark schemes rather than bring in their own paper versions.

2.1.9.7 E-marking schedule

The main marking phase took place over a period of two weeks, beginning Tuesday 29 May and ending Friday 8 June. A smaller amount of marking continued during 9, 10 and 11 June. Unfortunately, NCS Pearson did not manage to scan a sufficient number of scripts to keep the markers fully active during the first week. This was attributed to the conventional markers not returning a sufficient number of scripts on time (and, beyond that, to the schools not returning scripts to the conventional markers on time). Although more work was scanned in the second week, as more scripts arrived, some of the markers

had completed their 40 hour allocation and were unavailable for further marking. NCS Pearson also suggested that the e-markers were marking particularly quickly.[6]

The extent to which delays might have been attributable also to limitations in the implementation of the scanning technology is not clear. For example, the scanning was not being conducted at the rate that the machines were, in principle, capable of. On the other hand, NCS Pearson explained that they had scanned and made available everything that had been returned within the marking period.

As a consequence of the early delays, NCS Pearson was not able to process the full complement of scripts. Only around half of the intended volume was eventually marked (and a substantial number of scripts, that had been scanned but not marked, were left over at the end of the marking period – particularly for English). Thus, markers had insufficient work during the early stages and work was left incomplete at the end.

2.1.9.8 E-marker praxis

The expert markers chose to adopt somewhat different working patterns. Some worked for a smaller number of days but worked longer hours (i.e., four days at ten hours per day), while others worked for a larger number of days and worked shorter hours (i.e., ten days at four hours per day). In fact, markers had somewhat more flexibility than this and did not work precisely the same hours each day. Generally speaking, though, they all worked for a period of around 40 hours (which is what they were originally contracted to work). Most of the markers chose to mark between 7am and 6pm, although a smaller number chose to come in for evening shifts.

Senior markers worked considerably longer hours to ensure the highest standards of monitoring, administrative liaison and supervision. They were extremely approachable and appeared to be essential to the successful running of the operation.

Break times were not formally enforced for the e-markers and they took breaks as and when they wanted. In practice, breaks were often obligatory due to the lack of work available. During training, it was mentioned that, by law, a ten minute break should be

---

[6] It may well be true that the markers were marking faster than they would have done, had the tests assessed pupils across the full range of ability. The Y7PTs were taken by pupils of lower ability, meaning that many of the responses were easily identified as incorrect, or had simply been left blank.

taken for every hour spent looking at a screen (although NCS Pearson trainers pointed out that this pertained only to users who viewed screens continuously, that is, incessantly).

2.1.9.9 E-marker supervision

Originally, the intention was that the majority of the supervision might be undertaken remotely, via electronic messaging systems. However, in reality, the supervisors spent a lot of time in face-to-face interaction with the markers and resolved most of their queries in this manner.

Yet the Chief and Senior markers still had a formal supervisory role to undertake electronically – dealing with 'sticky noted' items (for clerical and response selection markers) and items discussed within (expert markers') messages. Importantly, the clerical and response selection sticky notes had to be resolved individually before each item could be processed through further stages. The electronic supervision of data entry, response selection and clerical markers was managed through the DWS Supervisor program. The electronic supervision of expert markers was managed through the Netgrade system.

2.1.9.10 E-marker adjudication

The Chief and Senior also had a formal role as adjudicators. When there was conflict between data input, or marks awarded, in relation to the same pupil response by two data entry, response selection or clerical markers then adjudication was necessary. The decision as to whether this was required was managed automatically through the DWS Post Edit program, and the adjudication process was managed through the DWS Adjudicator program. During adjudication, senior markers were presented with items for which two markers disagreed. They simply decided which of the two markers was correct and validated their value (or input the correct value if neither was appropriate). Once the adjudication had been completed, the result was taken as final.

For expert markers, adjudication was only undertaken as a result of one of the three automatic monitoring functions: between-marker reliability (see below). The other functions were deemed purely to serve a monitoring purpose. The adjudication of expert markers was managed through the Netgrade system, using the Scoring Client function. It presented only items for which marks provided by the two initial markers were in disagreement. Within Scoring Client a Scoring History option was available as a pop-up menu which showed the marks input by the initial markers. The adjudicator input the correct mark in Scoring Client.

2.1.9.11 E-marker monitoring

Only the expert markers were formally monitored for quality of marking (as items in the remaining marker categories were double marked). Four types of monitoring function were provided by NCS Pearson within the Netgrade system: between-marker reliability; within-marker reliability; absolute-marker reliability; and a manual monitoring function. The first three of these were automatic functions, while the final one could be selected as and when desired by the supervisor.

It is worth noting that, despite receiving conventional training and standardisation, the expert markers did not receive the normal 'S1' supervision that accompanies the early stages of conventional marking.

2.1.9.11.1 Between-marker reliability

This function, termed 'reliability' by NCS Pearson, was intended to determine the extent of agreement between a particular marker and other markers (who marked the same item clips). The way in which this was achieved seemed to be roughly as follows.[7]

From the original pool of all item image clips to be marked by expert markers (i.e., all pupil responses to all expert questions) a fixed proportion were selected. This proportion was chosen to be 10%. Each of these chosen item image clips was replicated and placed back in the pool with their 'twins'. Markers who logged on to each question were then allocated item images in a relatively random fashion, with the proviso that a single marker could not be allocated the same item image clip twice (with the exception of those formally generated for within-marker reliability). Thus, by the end of the marking period, 10% of items would have been marked twice, by different markers.

Two significant points are worth mentioning at this stage. First, under this procedure, the fixed proportion of items (10%) does not apply to individual markers, or to individual questions, but to the entire item pool. Thus, each marker may have a different proportion of items re-marked for each question that s/he marks. Furthermore, when aggregated

---

[7] It is not clear the extent to which the following explanation is approximately literal or more figurative. The algorithms were determined by programmers in China and communication received from them, via NCS Pearson, was not entirely transparent. However, the explanation seems to convey the gist of what must have happened.

across questions, different markers may have a different proportion of items re-marked. (Appendix 5.3 reveals the extent of these differences.)

Second, every item that is marked twice provides monitoring information on two markers. Therefore, if the proportion of items selected for double-marking is 10%, then a significantly higher proportion of each individual marker's allocation will be re-marked (in fact, closer to 20%).

2.1.9.11.2 Within-marker reliability

The second function, termed 'consistency' by NCS Pearson, was intended to determine the extent of agreement between marks awarded by a single marker to the same item on two separate occasions. The logic of this function was to replicate a fixed proportion of item image clips allocated to a particular marker and to re-allocate them to the same marker at a later stage.

Although NCS Pearson initially reported to the Evaluation team that the fixed proportion would be 10%, it eventually turned out to be closer to 3% or 4%. However, this monitoring function was not part of the original specification, and the lower sampling frequency did not constitute a failure to meet an agreed target.

2.1.9.11.3 Absolute-marker reliability

A third function, termed 'validity' by NCS Pearson, was intended to determine the extent to which each individual marker was in agreement with the 'absolute' standard set by the Chief Marker. Thus, for each expert question, a specified number of item image clips were marked in advance by the Chief Marker.[8] At set periods – say every three-quarters of an hour – each marker was to have been fed one of these 'seed' items. As data accrued, this would have constructed a picture of the extent to which each marker was in step with the Chief, for each question.

Owing to a problem in loading the Chief Marker's marks, this function failed to function. This was particularly unfortunate because it was, by far, the most important component of the monitoring process.

---

[8] Indeed, these responses could equally have come from the pre-test as from the live administration and could, therefore, have been incorporated into the system well before the marking began.

2.1.9.11.4 Manual monitoring function

Finally, supervisors were also able to view items that markers had marked manually. This enabled them to investigate specific concerns that they might have with the marking of particular markers on particular questions.

This manual supervisory monitoring was achieved through the Check Scoring function of Netgrade. Supervisors were unable directly to amend the marks that they observed, as this could only be achieved at a later stage through the Scoring Client function.

2.1.9.11.5 Using the monitoring data

Statistics for each of the monitoring functions, computed daily by question for each marker, were available on the NCS Pearson intranet. However, the senior markers were not aware of this data, or at least of how to access it, until the second week. At this point it was understood still to be "in its development stage". The data often took five or six minutes to download from the intranet.

On the basis of their analysis of marker performance, supervisors were empowered to debar markers from continuing marking, or from continuing with specific questions. This action was never taken during the Pilot.

It should be noted that the NCS Pearson software systems also generated a wealth of additional data for monitoring markers for reasons other than quality of marking (i.e., monitoring marking speed) and for monitoring the system more generally (i.e., monitoring the processing of batches). These functions, and data arising from them, are considered extensively in Section 5.

2.1.10 Data processing and storage

Final marks for each item of each paper were exported to pupil databases, in a process managed through the DWS Transvertor program. Marks were aggregated within and across papers and level thresholds applied.

It should be noted that, unlike in the conventional marking process, scripts were not borderlined.

2.1.11 Results processing and reporting

The system for reporting upon results ensured that pupil and school profiles were produced alongside LEA and national data. Although not formally carried through for

2001, the system was equipped to despatch reports on school and pupil performance to schools in DfES format (using the DfEE Extract Program).

Similarly, although not formally carried through for 2001, the system was equipped to produce CD ROMs for schools presenting pupil script images (using the CD ROM Publishing program).

Trial CD-ROMs were produced for a sample of schools, but not distributed. In addition, a 'generic' CD-ROM was produced for illustrative purposes and sent to a sample of schools for review. However, it was not within the remit of the Evaluation to consider this aspect of the Pilot.

2.2 Strengths and weaknesses of the Pilot

The Pilot was successful in procedural terms. NCS Pearson demonstrated (at least on a small scale) that the marking of UK national tests can be undertaken using a centre-based model of on-line marking.

Perhaps inevitably, though, there were certain aspects of the implementation that were subject to misunderstanding, that encountered unanticipated obstacles, or that simply did not work. The following sub-sections, largely replicating the sub-sections of 2.1, will focus upon these issues with a view to lessons that may be learned for future years.

2.2.1 Answer booklet modification

In their original submission to QCA, NCS Pearson outlined numerous ways in which answer booklets could be modified to enable a far more effective use of the new technologies. In particular, alterations were proposed to the response formats of certain questions, that would facilitate the automatic marking of pupils' responses. Concerned that this might significantly alter the demands placed upon pupils in the Pilot, QCA required that NCS Pearson should not make any such modifications. Therefore, only modifications that were required to enable the scanning of scripts were implemented.

Clearly, the automatic marking of certain questions would promise significant reductions in marking costs. However, NFER concurs with QCA that even apparently trivial modifications to test papers can change the functioning of test items. If pupils were to be presented with novel response formats this would be likely to introduce an element of 'construct-irrelevant variance' into the resultant scores (i.e., some pupils would fail to gain marks that they actually deserved simply because they failed to appreciate exactly how to respond). The USA has a long tradition of objective testing with constrained

response formats, meaning that teachers and pupils are used to the kind of innovation suggested by NCS Pearson. The UK, on the other hand, has no such tradition (particularly for younger age groups) and an innovation as apparently trivial as response format alteration would therefore need to be managed very carefully.

NCS Pearson also encountered a problem that led to a significant number of scanning rejections: pupils scribbling, or writing, over the ID and timing marks added to the answer booklets.[9] Once again, pupils in the UK are not used to responding to tests in ways that are optimal for the implementation of new technologies. The significance of not obscuring barcodes was not apparent to many of them (although they were not actually instructed not to scribble over them). Such problems suggest either that the culture of testing in UK schools will have to change to accommodate technological innovations or that the new technologies will have to adapt to accommodate the UK testing culture.

2.2.2 Software preparation

At the outset of the Evaluation it became clear that there had been some misunderstanding between NCS Pearson and QCA concerning the central software of the Pilot. Whereas QCA understood that the ePEN system would manage expert marking, NCS Pearson had decided that this function would be provided by Netgrade.[10]

Despite the fact that Netgrade and ePEN perform similar functions, the Evaluation team felt that the use of Netgrade rather than ePEN was not an insignificant difference, particularly as NCS Pearson propose to use ePEN in future projects. Had the project been a formal trial, rather than a pilot, this would have constituted a flaw. Even as a pilot, though, it was an unfortunate shortcoming. Clearly, the Pilot could provide no information on how a delivery via ePEN might have differed, on whether it might have been more or less effective, or whether it might have resulted in entirely new kinds of problem.

---

[9] The minimum rejection rate was 1.2% for maths arithmetic and the maximum was 5.1% for English writing.

[10] NCS Pearson explained that they group all of their on-line marking solutions under the branding of ePEN in an attempt to avoid confusion. As Netgrade and ePEN perform similar functions, it was not felt appropriate to make specific distinctions.

The Netgrade software was provided by the Chinese subsidiary of NCS Pearson (CES) and employees from China were seconded to integrate the system within DWS. As will be discussed below, Netgrade did not always deliver effectively and it was not always clear to the UK software engineers exactly why.

2.2.3 Script despatch to schools

The personalisation of scripts for pupils worked effectively in the main. However, advance personalisation will always have to deal with pupils who leave schools, and who join schools, within the months between pupil data collection and test administration.[11] A number of schools photocopied scripts to resolve such problems (which defeated the objective of script personalisation). Once again, the introduction of new technologies will change the culture of testing in UK schools and this change will need to be managed effectively. Alternatively, the new technologies will need to be adapted to accommodate the idiosyncrasies associated with the pre-existing culture.

Clearly, in some instances, the adoption of new technologies will force a change of testing culture and it will be imperative that all teachers are brought up to speed. For example, it will be crucial to ensure that all teachers understand the necessity that pupils receive the answer booklet that has been personalised specifically for them. One school had apparently given out answer booklets without regard for names.

This raises the issue of exactly what quality assurance procedures were in place to ensure that pupils actually had answered in the correct booklets. It was not clear to the NFER what procedures – if any – had been put in place by NCS Pearson (nor what procedures would be in place in future years). Although one of the work instruction sheets (QCA Editor Names) detailed procedures for dealing with pupil names, it was not clear that this was intended as a quality control procedure on receipt of scripts. Nor was such a procedure detailed on the (Uncontrolled) Quality Plan for the Pilot. For future years, it would seem important that some kind of formal validation of pupil identity – a verification of printed name against written name – be carried out upon receipt of pupils' scripts.

---

[11] In addition, the Pilot faced problems specific to the Y7PT, particularly those resulting from schools being unsure of the criteria for selecting pupils to participate in the test (e.g., those registering all of their pupils).

2.2.4 Test administration and conventional marking

Few complications were noted in relation to test administration and conventional marking (with the exception of those mentioned in 2.2.3).

2.2.5 Script return to NCS Pearson

The script despatch and return involved processing a very large amount of paper. In the main this was handled well by a workforce that was predominantly composed of temporary staff. Occasionally a batch was temporarily mislaid, but these all turned up eventually.

Upon scale-up, the massive increase in volume of paper involved – particularly the physical capacity for scripts to be processed, transported and stored effectively – will be an issue for deliberation.

A further example of testing culture that raised problems for NCS Pearson concerned the non-return of personalised scripts. The automatic reconciliation during script return threw up a significant number of queries that had to be dealt with before processing those batches further. Dealing with a query required personally contacting a school to determine why a particular script had not been returned. This slowed the processing down. Moreover, some schools had already finished for the half-term vacation and resolution was not possible. It seemed likely that the most common cause was a script not having been returned due to a pupil having left the school prior to, or being absent on the day of, the test administration. Future contractors will need to consider how to address this problem.

2.2.6 Script scanning

2.2.6.1 Guillotining

Two of the administrative staff had been trained on the guillotine machines and seemed extremely proficient. Throughout the observation, no serious problems were observed with the guillotining, and it seemed unlikely that this process would lead to substantial bottlenecks in a scaled-up system unless the machines malfunctioned and could not be repaired or replaced.

2.2.6.2 Scan Master

The scanning of scripts was affected by a number of problems, many of which have already been mentioned. One anticipated obstacle that became apparent at the outset was that pupils in the UK are allowed to respond to the tests using a range of writing implements, including pencils of different hardness and pens of different colour. Typically, in the USA, pupils are permitted only to respond using pencils of a specific type. As such, the scanning machines had been optimised to record annotations written in different colours and substances. In fact, some thirty or forty scripts had been written in red pen and thus were not readable after scanning. These scripts had to be copied in different ink before they could be re-scanned and marked.

It has already been noted that some pupils interfered with the timing or ID marks pre-printed on the answer booklets. This led to rejects not only when the marks could not be read, but also when interference meant the marks were read in the wrong way (i.e., when a pupil was not identified or identified as another pupil).

On occasion, the scanning machines were responsible for delays, owing to technical failure (for example, on the afternoon of 7 June, one of the scanners was out of use for a period of about an hour due to problems with the reader component). Such problems required specialists in network development for their rectification. Machine failure would seem to constitute one of the major threats to the large-scale implementation of new marking technology.

2.2.6.3 Scan Master Exceptions

The number of 'exceptions' which had to be scanned by hand was considerable; it slowed the turnaround time, as each batch waited for its exceptions to be hand-scanned before it could continue through to re-collation (presumably to reduce chances of mislaid scripts). It should be noted that this process was done entirely by hand, with pages manually positioned on the scanner rather than being fed in automatically; hence, hand-scanning was a very labour intensive process.

Problems encountered by expert markers, that might well have been attributable to error in the hand-scanning process; included:

1. pages of writing scripts presented upside down;

2. pages of writing scripts presented in the wrong order.[12]

On occasion, clips were encountered that were entirely unreadable. It is not clear whether NCS Pearson necessarily ensured that such problems were rectified for all pupils before the marking was completed (recognising the fact that this was merely a pilot). Of course, it would be essential that procedures were in place for dealing with such issues if they arose during a formal testing context. However, addressing such problems might not be trivial. If the system had failed successfully to scan a script – and this was only identified once the script had been returned to a school – there might be significant problems in reclaiming it for a second scan (particularly if the school had already broken up for the vacation).

2.2.7 Script despatch to markers

The script despatch to markers was successful. The only problem that was noted concerned a batch that had been returned to a school without having been scanned. It appeared that the DWS software could not identify whether a batch had been scanned or not from its batch header. The errant batch was traced and its return requested.

2.2.8 Item image administration

2.2.8.1 Image capture procedure

On 1 June, it was reported that the first 134 batches scanned had not stored a full-page image. This appeared to be due to a misunderstanding concerning procedural requirements. These batches (2,714 scripts) were included with subsequent scanning work, further slowing the flow of item image clips into the distribution server. The majority of these early scripts missed the four day turnaround target.

2.2.8.2 Image distribution errors

Some of the items seemed to have been stored incorrectly, and this became noticeable at the distribution stage. This occurred on several occasions for writing scripts, with *Mystery Solved* and *Three Wishes* texts being allocated to those who were marking *Tried*

---

[12] This was sometimes due to children starting on spare paper before continuing in the booklet. NCS Pearson explained that the software could be revised for future years to ensure that this kind of problem did not occur again.

and Tested. The explanation appeared to have been that some pupils omitted to mark the box on the front of the booklet to indicate their choice of writing response. In this event, the DWS software identified a writing response as *Tried and Tested* by default.

Other, similar problems included one English marker receiving an entire answer booklet (on several occasions). As mentioned earlier, a number of scripts were received upside down or in the wrong order.

2.2.8.3 Repeat allocation

On 7 June, one of the maths markers received the same item as many as five times. The Netgrade developer who was consulted about this problem was unsure as to its cause, although she suspected that it may have been due to the generation of monitoring clips by the software: it seemed that the system had spawned a larger number of these 'twin' clips than expected (and this may have been related to the marker in question not having marked for a few days). Exactly why this might have happened was not at all clear.

The impression gained during observation was that 'computer faults' akin to the repeat allocation problem occurred more frequently than was reported, as markers tended not to be interested in their cause and generally ignored them. Their exact prevalence was not clear.

2.2.9 Electronic marking

The marking software generally appeared to achieve the basic marking function well. The markers were able to navigate the software although it was suggested that the system was sometimes more 'cumbersome' to use than it might have been, particularly for administrative and supervisory functions.

One small problem for data entry was that at least one paper marker, contrary to her instructions, had marked correct spellings with a tick. This looked like an oblique to those doing data entry.

2.2.9.1 E-marker structure

The allocation of different item types to different markers functioned well and appeared to be an appropriate way in which to maximise the potential of the human resources available.

Likewise, the classification of items according to marker type was effective. The only change in classification occurred for the spelling items. These were initially allocated to response selection markers and were subsequently designated as clerical items.

2.2.9.1.1 Response selection markers

No specific issues were noted in relation to response selection markers.

2.2.9.1.2 Clerical markers

The Chief Marker for English suggested that spelling should not be marked by clerical markers. She felt that it required more skill and insight into children's writing than the clerical markers possessed and that they were marking this paper too quickly.

Whereas response selection markers input responses as they appeared on scripts, clerical markers input marks directly. However, despite spelling having been (re-)assigned to clerical markers, both the supervisor and adjudicator functions allowed the actual spelling response to be input (rather than the mark). It appeared that this had caused some problems for networking staff, who occasionally had to reinterpret results from DWS Post Edit.

2.2.9.1.3 Expert markers

2.2.9.1.3.1 Speed of marking

The processing of items for expert markers was most affected by the availability of clips to mark. As markers were allocated to no more than two questions at any one time they frequently ran out of clips and were presented with a 'No Task' screen.

If there were no to-be-marked clips for any of the allocated questions, then a marker could ask to be allocated to a different question. The reallocation process was quite lengthy and complex, and required the marker to log off. Running out of work was the greatest source of frustration for the expert markers, who were keen to get on with the job and sometimes found it hard to find a senior marker to reallocate them. It was also the greatest source of frustration for the senior markers, who found the reallocation process complex and distracting. Indeed, as work was only coming through as a 'trickle', a lot of time that they had hoped to use in supervision and adjudication was taken up in reallocating.

2.2.9.1.3.2 Marking accuracy

Expert marking required the marker to carry out two steps for each clip:

1. to select the option button that corresponded to the mark they wished to award;

2. to confirm that selection by clicking on a 'Commit' box.

Both of these were done with a mouse and the area to be clicked was small in the mark selection case. Once markers had become experienced they selected and confirmed quite rapidly. This introduced a risk of confirming the wrong selection. No information was gathered on the frequency with which such errors were committed, yet it seems likely that some (or even many) may have been.

In fact, the Netgrade system defaulted to an acceptable value ('no response'), without requiring that a marker complete step 1 at all. Thus, if a marker mis-clicked, failing to select any of the option boxes, then – not having noticed – s/he could still Commit the default and would be taken directly to the next clip.

In future pilots it will be important to explore the extent to which such errors may occur. This is particularly crucial when single (rather than double) marking is employed for expert markers – as errors will not be picked up through adjudication. It will also be important to develop software, or procedural adaptations, that may eliminate the possibility of such errors.

2.2.9.1.3.3 Messaging

To note problems or queries, expert markers were required to send messages using the Netgrade message function. This required copying and pasting an item's Encrypt ID and Item No. into the message text (as it was not included automatically). Some of the expert markers that lacked training and experience with computers were unaware that this information had to be included. Others were aware that it was to be included, but were unaware of the procedures to copy and paste and noted it on paper before transferring it between windows.

The messaging system faced additional problems, as the Chief Markers found they did not have enough time to answer written queries, and thus feedback did not occur. Moreover, to receive messages relating to individual questions from specific papers, a Chief Marker had to log on to that particular question (using a different password for each one reviewed). In the case of English, this meant having 26 different log on names and

passwords. Senior markers were obviously frustrated by this procedure. This was compounded by the fact that there was no indication that they had waiting messages relating to any of the items; that is, they had to log on to each item before finding out whether or not there were any messages to respond to. In short, the messaging system for Netgrade was very poor. In practice, most guidance, feedback and querying was conducted in person by the senior markers, on the spot, in response to raised hands. This meant that some queries were never addressed. Clearly, the system could not have dealt effectively with on-line supervision under a remote web-based model – even in a centre-based model it would require a high ratio of supervisors to markers.

2.2.9.1.3.4 Guessing

There was no option for expert markers to 'pass' on an item and forward it directly to a supervisor. To enable them to be taken to the next item they were required to provide a mark for the previous one. As such, and due to the ambiguity of supervision procedures, there would have been a natural temptation for markers to guess when unsure and hope that any errors would be 'picked up by the system'. In fact, as only a small minority of clips were double-marked, the system would most likely not have 'picked up' mistakes unless an additional message had been sent (and, in practice, unlikely anyway because the message system largely failed). Particularly if single marking is to be employed in the future, it would seem advisable to allow even expert markers to pass on problematic clips, and to send them directly to a supervisor.

2.2.9.2 E-marker appointment

The major problem encountered in staffing the Pilot was the recruitment of an insufficient number of sufficiently skilled markers. Even the markers that were recruited did not have the requisite experience of conventional marking. However, they took to the Pilot well and appeared comfortable with the task demands.

In the middle of the second week, a decision was made to train some clerical markers to adjudicate and supervise data entry and response selection markers. This was to reduce the backlog of those scripts that were held up at this stage.

The newly trained adjudicators were often required to resolve fairly straightforward discrepancies (for example, '80000' versus '80,000' or '80 000'). On other occasions, the

discrepancies were substantially more difficult to adjudicate.[13] They were advised to refer items that they were unsure about to the Chief Marker. However, there was evidence that the gradation in task difficulty tempted novices to apply their own judgement more than perhaps they ought. Moreover, as there was no function within the DWS Adjudication software to 'pass' on an item, this might have tempted them further when the senior marker was not immediately present.

2.2.9.3 E-marker training

The quality of the training in the software at NCS Pearson was perhaps not as thorough as it might have been. There was clearly a lot to be learned by employees who were not necessarily even familiar with computers.

Although User Guides were provided, it seems likely that these were of much greater benefit to those who already had a sound grasp of the task and of the software procedures involved. The User Guides might have included more basic detail on specific shortcuts (e.g., to avoid logging on and off to look for new tasks) and general procedures (e.g., copy-paste functions).

It is important to realise that the expert marker monitoring functions included within the Netgrade software did not compensate for the omission of the crucial quality assurance stages represented by the conventional first, second and final samples. Successful completion of the first sample (which follows successful completion of the training and standardisation samples) is, effectively, a licence to continue with the live scripts; that is, the first sample is a test of whether the mark scheme has been sufficiently appropriated by a marker and must be approved before any live marks are formally submitted. In contrast, there was no such quality assurance test before the expert e-markers began submitting live marks for individual questions. Even if the 'validity' monitoring function had worked, it would only have provided information that a marker was below standard long after s/he

[13] An example of an item that was difficult to judge is one that had two numbers, one written on top of the other. On paper, this might have been possible to judge; on a screen it became very difficult. Incidentally, NCS Pearson explained that one solution to such a problem would be to enhance the definition during scanning (to capture in 256 levels of greyscale rather than 16).

had begun submitting live marks. This is an oversight that ought somehow to be addressed in future years.[14]

Unfortunately, the incorporation of an on-line procedure comparable to the conventional first sample might require smaller marker-to-supervisor ratios if formative feedback were to be provided immediately to markers working in a marking centre. Note that this feedback would have to be provided immediately by supervising markers, in contrast to the conventional situation which is paced by the postal service and by telephone accessibility. This might be particularly problematic if all markers were to commence at the same time and, therefore, require feedback simultaneously.

As, during the Pilot, there were no new technology-mediated exchanges between e-markers and supervisors comparable to the first sample stage, it is not clear whether or not this kind of support could have been achieved effectively (or whether a more personal touch would have been necessary).

It should be stressed that the adoption of new marking technologies could, conceivably, force very significant changes in both the culture and the structure of marker training and supervision. The conventional broad-based triangular hierarchies – markers below team leaders, below senior markers, etc. – may turn out to be inappropriate for new technology marking, which might instead force flatter structures. Before the new technologies are implemented on a large scale it will be essential to consider how the marker training and supervision structures will need to change and how this can best be managed. It might, for instance, require either large increases in the number of markers of team leader status, or large decreases; both of these changes would require sensitive and skilful management to implement successfully. If structural change is considered likely, it would be sensible to pilot new approaches sooner rather than latter as the approach adopted may be a major determinant of the success of the entire enterprise.

---

[14] As part of the Pilot training, NCS Pearson ensured that additional training and standardisation scripts were completed by the expert markers to ensure that they were marking at an appropriate standard. In future years it would clearly be possible to employ similar conventional procedures for ensuring that markers were sufficiently qualified. However, it is anticipated that QCA would expect the training, standardisation and qualification of on-line markers in future projects to be more automated in line with the new technology procedures employed.

2.2.9.4 E-marker working environment

The e-marker working environment was generally felt to be very satisfactory. The supervising markers no doubt contributed to this impression by making an effort to attend to markers' concerns personally and at their workstations. The Chief Marker for English made one recommendation which, presumably, related to ease of supervision: that clerical markers should work in the same area as the expert markers.

2.2.9.5 E-marker administration

The administration of the Pilot – which included the organisation of staff rotas, responding to technical queries, etc. – was generally very successful. However, during the second week of the pilot, administrative support was less in evidence, as staff returned to desks in different areas of the NCS Pearson HQ and may have appeared less accessible.

2.2.9.6 E-marking software

The software functioned largely as intended. However, there were certain 'glitches' that raised concern and that are worth highlighting. Some of these were design issues, that frustrated many markers throughout the entire process, but that were probably amenable to simple adjustment and therefore would not have serious implications for future development. Others seemed to be equally pervasive but, perhaps, less straightforward to rectify and therefore more of a problem. Yet others were infrequent and, owing to their oddity, of unknown significance.

2.2.9.6.1 Design obstacles

The tiresome requirement for markers and supervisors to log onto individual questions to determine whether they had any items to deal with was an obstacle amenable to a small amount of programming. Other problems also seemed to fall into this category. For example, one of the senior markers for English felt that a useful improvement to the Netgrade software would be the ability to allocate all suitable items to a marker at once, and then let them choose their own work from this pool, perhaps with a counter which indicated how many of that item had been done, and remained to be done. The extant system meant that only one item from a particular paper (e.g. handwriting, reading or writing) could be allocated at a time, making the necessity for reallocation much more frequent.

To provide an even more cumbersome example, consider the procedure required for a senior marker to check the marking of an expert marker. First, it was necessary to log in

as an administrator and move the marker to the relevant question group (in the same way as allocating a question to them). The senior marker then had to log out, and log back in as a supervisor, and log in again to that question group. Moreover, this process could not be undertaken while the marker was present and using the system.

In addition to problems that seemed amenable to a simple solution there were others that appeared somewhat more intrinsic. In particular, in DWS, if the first marker to mark a particular item clip had sent it directly to a supervisor (with a sticky note) then it could not be marked by a second marker until the supervisor had dealt with it. This led to some work flow problems as the number of clips to be dealt with by supervisors built up.

2.2.9.6.2 More mysterious problems

There were other, more mysterious, problems. On 8 June, using the Scoring Client/Scoring History function on Netgrade (to investigate those handwriting scripts with discrepant marks), one of the senior markers found that, for at least one of the apparent discrepancies, the same mark had actually been awarded. Further, he found at least one discrepancy where he was named as one of the markers, despite never having marked any of these scripts. The software developers were unable to explain either of these occurrences. Another minor mysterious problem involved messages sent by maths markers which were received by English senior markers.

The process of on-line supervision of expert markers also raised problems when initial marks needed amending. The Netgrade software did not allow this within the supervision function and clips could only be 'flagged'. This became a somewhat mysterious procedure and, by the end of the marking, none of the senior markers seemed to know how to access 'flagged' items, exactly where they had gone, nor whether they were eventually resubmitted to markers, supervisors, or to no one at all.

The mysteriousness of certain Netgrade functions was compounded by the English prompts and labels having been quite badly translated (from Chinese), which made them difficult to understand.

2.2.9.6.3 On-line mark schemes

Most of the maths expert markers were prepared to use the on-line mark schemes (when required). On the other hand, many of the English expert markers expressed a preference for their own paper mark schemes.

There are good reasons to question whether the use of the on-line mark schemes was appropriate (let alone straightforward). During the national training process, markers are instructed to annotate their mark schemes and to refer to these annotated mark schemes when marking. It was not possible to annotate the on-line mark schemes.

Moreover, the on-line mark schemes were not very convenient to use. To view the clips and mark schemes simultaneously required windows to be re-sized; alternatively, switching between screens was possible, but not simple for those not used to 'toggle' technology. Viewing the on-line mark scheme appeared to be more of a problem for the handwriting and writing tasks, as it appeared that a response could not be scrolled while the mark scheme was being consulted.

2.2.9.6.4 Unanticipated validity threats

Finally, it is worth noting the danger that even apparently beneficial aspects of the software could have unanticipated adverse impacts upon the quality of marking. The ability to 'zoom' is a potential case in point. Thus, if different markers were to view handwriting responses at different magnifications then this would probably have an impact on the nature of the judgemental processes involved and could, conceivably, introduce additional construct-irrelevant variance into the marks awarded. Likewise, marking spelling responses at high magnification might alter the judgemental task significantly.

2.2.9.7 E-marking schedule

Apart from problems arising from the lack of items to be marked in the early stages, the marking schedule seemed to work well, with markers making use of the potential for working flexibly. Further issues concerning the schedule are discussed in subsequent sections, particularly in relation to markers' attitudes.

2.2.9.8 E-marker praxis

On-line marking is a new way of working and would change the culture of marking in many ways. As such, it was important to consider how well the markers adapted to the new practices and working environment during the Pilot. It should, however, be noted that the expert markers could not be considered a representative sample of conventional markers and it would be unwise to extrapolate from their responses to those of the marking community in general.

Generally speaking, the expert markers had a positive attitude to their work, even during frustrating periods such as occurred when levels of work were less than expected and task re-allocation was frequent. While this positive attitude may to some extent have been due to the novelty of computer marking, or to the potential for social interaction with other teachers, it nevertheless seemed significant. The markers generally appeared to be diligent, with a high level of precision and a commitment to producing accurate work.

The clerical and response selection (data entry) markers had quite repetitive and potentially boring jobs. In general, vacationing students carried out this work, and – despite the potential tediousness of their tasks – seemed capable and focused.

It is inevitable that some cultural changes will be harder to implement than others. One useful example of this resulted from the necessity for expert markers to log out of Netgrade when they were not actually marking items (including when they had no tasks available to mark). It appeared that few of the e-markers actually did this consistently. The knock-on effect was that the automatically computed indices of marker efficiency and marking rates were neither entirely reliable nor valid. As such, these indices were not relied upon and productivity was computed solely on the basis of the amount of time for which item clips were viewed. The only problem with this computation occurred when markers had taken impromptu breaks, following presentation of a 'No Task' window, and had not returned before the next item had been delivered. As soon as the new item was delivered the clock would have started ticking, regardless of whether the marker had been present.[15]

2.2.9.9 E-marker supervision

While the marking supervisors were very diligent and supportive, it appeared that their role had not been fully thought through and that the procedures of e-marking were not necessarily supportive of the kind of supervision that they expected to provide.

Although it was originally intended that supervision be undertaken predominantly on-line, it soon became clear that a personal touch was required. (This would clearly have implications for the success of a centre-based system, but the implications would be more serious still for a remote web-based system.) Unfortunately, as much of their time was

---

[15] In addition, not logging off when taking breaks would constitute a potential security risk for markers marking remotely in a web-based system.

spent at other markers' terminals this led to a backlog in items that were awaiting a supervisor's attention.

As explained earlier, when supervisors did get to consider sticky note messages that had been sent to them by clerical and response selection markers, it was not clear whether their task was simply to award the appropriate mark or also to provide formative feedback to markers (who may have sent the item some hours previously). Thus, it was not clear whether the supervisors' on-line role was intended to be primarily one of quality control (noting and amending incorrect answers) or quality assurance (providing direct formative feedback to markers). Nor, when scrutinising items that had already been marked by expert markers, was it clear how they ought to amend those that they considered incorrect.

2.2.9.10 E-marker adjudication

The adjudication function appeared to be effective in the DWS system, although somewhat more cumbersome in the Netgrade system. However, the rationale for adjudication in the Netgrade system seemed somewhat under-theorised. NCS Pearson explained that they generally adopt e-marking models in which *all* items are double marked (whether response selection, clerical or expert). Any items for which the marks from two markers are discrepant are then sent automatically for adjudication. This is exactly what happened in the DWS system. However, as the vast majority of items in the Netgrade system were marked by only one expert marker, the situation was more complicated.

In the Netgrade system, only 10% of expert items were marked twice by different markers (for 'between-marker reliability') and only discrepancies arising from these items were sent for adjudication.[16] This meant that there was no such quality control for the remaining 90% of items (for which manual monitoring was intended to suffice).

There seemed to been no explicit decision concerning whether the double marking of expert items was intended primarily: to serve a **remedial quality control** function (to correct errors before the final product was released); or to provide an **accountability**

---

[16] Adjudication was not implemented for those items awarded different marks from the 3% to 4% of items that were marked twice by the same marker (for 'within-marker reliability'). This function was used only for monitoring purposes and the marks awarded to the re-mark clips did not count towards a pupil's mark total.

quality control function (to locate and quantify the degree of error within the system); or to support a **formative quality assurance** function (to help markers identify and overcome their mistakes and misunderstandings, during the marking process, to prevent further errors).

While the adjudication functions appeared effective, there were a number of relatively small glitches that are worth commenting upon. One irritation with the DWS system was the fact that adjudicators lacked immediate information on the component from which the discrepancy arose and on whether the question was clerical or response selection (only the question number was provided). The Netgrade system, likewise, lacked immediate information on the component from which the discrepancy arose.

A related problem was that, when selecting the full-page image of a discrepantly marked item, the DWS page images were numbered as two greater than the pages on the paper tests (that is, to select page 3, one had to request page 5).

2.2.9.11 E-marker monitoring

As noted above, there was an inherent lack of clarity over the rationale for the double marking of expert items: was it to serve a remedial quality control function, or to provide an accountability quality control function, or to support a formative quality assurance function? In a similar way, there was a lack of clarity over the way in which monitoring information was to be used. Once again, this may, to some extent, reflect a relative lack of attention to the specific problems that must be addressed when single marking is employed.

With double marking, remedial quality control is implemented for each item and (following adjudication by senior markers) the final product can be assumed to be of a high standard for all pupils.[17] As such, it is easy to understand why there might be less emphasis upon formative quality assurance. Indeed, in principle, monitoring data need only be adopted for accountability purposes: either during the marking, to debar clearly

---

[17] Note that double marking does not entirely eliminate marking error. Indeed, the larger the number of discrepancies observed, the larger the number of 'false agreements' (i.e., markers who agree on an incorrect mark) there will be. The probability of false agreement is likely to increase as the mark available for a question decreases (being at its highest for single mark items).

sub-standard markers from continuing, or after the marking, as a criterion for re-appointment or payment.

However, with single marking, there is no inevitable remedial quality control. Moreover, when remedial quality control is introduced through the double marking of samples of items, it is often introduced to support a formative quality assurance function through the identification of sub-standard marking and markers. With these comments in mind, we can turn to the four expert marker monitoring functions.

2.2.9.11.1 Between-marker reliability

When the initial measurement data feeds were received from NCS Pearson (see Section 5) there was some confusion over the contents of the data files, and two major apparent anomalies were observed for the between-marker reliability monitoring data:

1. Across each of the six papers, for each individual question, the ratio of items double marked to all items marked varied substantially across markers – that is, while one marker may have had 2% of all items that she marked (for, say, English reading Q4) re-marked, another marker may have had 98% of her items (for English reading Q4) re-marked.

2. Across each of the six papers, when roughly averaged across questions, the ratio of items re-marked to all items marked was roughly 20% for each marker. This contrasted with the supposed sampling frequency of 10%. That is, each marker had around 20% of all items marked (for, say, English reading) re-marked. Again there was some fluctuation between markers, although not as extreme as at the individual question level.

In fact, it became apparent that these were not actually anomalies in the data. However, what these characteristics did suggest was that the algorithm underlying the between-marker monitoring function was not ideally suited for providing the kind of formative quality assurance data that might be required.

First, if a supervisor is to be reassured that each marker is marking each question successfully, then it will not be helpful if 98% of a question allocation is re-marked for one marker while only 2% of another marker's allocation for the same question is re-marked; the supervisor will have an unnecessarily accurate picture for the former and an inaccurate picture for the second. Instead, the supervisor needs a similar proportion of items re-marked for each marker.

Second, if it is intended that around 10% of a marker's allocation be re-marked, then it would only be necessary to ensure that just over 5% of items were double marked (as each double marked item provides information on two markers). The algorithm which was designed to select 10% of items seems to betray an implication that the monitoring data were intended to be used more for accountability quality control (i.e., audit) rather than for formative quality assurance.

Finally, there is a general question of whether random monitoring throughout the marking period is the optimal approach, assuming that the most important aim is to provide information for formative feedback. What happens during the conventional marking process is that samples are routinely assessed right at the outset (the first sample) and at around two-thirds of the way into the marking (the final sample). If single marking continues, with the requirement that monitoring data be provided primarily for formative quality assurance, then to focus that monitoring right at the outset (and, perhaps, later on as well) would appear the more appropriate option. However, instead of being on a script level basis, the monitoring would be on a question level basis. Moreover, there would be no reason why not to exploit some form of double (or better, triple, quadruple, etc.) marking system rather than the conventional system whereby a team leader re-marks a different set of sample scripts for each marker.

Of course, this raises a very important question. Would it actually be more effective (reliable, valid and manageable) simply to employ 100% double marking even for expert items? It seems likely that there would be a payoff in terms of marking reliability (and, hence, validity). However, there might also be unanticipated payoffs in terms of paying less attention to the provision of formative feedback and, thereby, freeing the more senior markers for the crucial task of adjudication.[18] Indeed, the initial sampling and formative feedback stage could be largely automated – for delivery at the question level – through the construction of training and 'certification' modules based upon carefully selected seed items. On the other hand, there might be significant costs, for example, if expert markers felt that their professional autonomy was being reduced (by double marking) then they might be less willing to participate; alternatively, they might exercise less care, leading to an increased number of items needing adjudication, which might over-load the more senior marking teams. The extent to which any of these possible effects might be

[18] Moreover, the adjudicators would only have to be good markers and would not require additional skills of team leading and supervision (once more optimising the available human resources).

attenuated or exaggerated by linking payment (or future employment) to specific accuracy criteria is not clear and would need to be considered.

The decision whether to remain with single marking for expert items or to employ double marking across all items types is a crucial one and one that should be considered for any future project. In many ways, the choice will shape the future of new technology marking and may be a major factor in the extent to which it is successful.

2.2.9.11.2 Within-marker reliability

As for between-marker reliability, the percentages of items sampled varied across markers, which made useful comparison more awkward and which betrayed that the function was probably not designed with the provision of formative feedback in mind. Indeed, it was not at all clear what the intended use of the within-marker reliability function was. This seemed to be an example of data being provided without an explicit formulation of the questions that they were supposed to answer.

2.2.9.11.3 Absolute-marker reliability

The absolute-marker function failed. Once again, the Evaluation team was not provided with explicit evidence concerning how the resultant data would have been used, had the function been successful. The data on absolute-marker reliability would only have accrued slowly (based upon a periodical sampling algorithm) and, by the time sufficient data had accrued to evaluate a marker's quality for a particular question, s/he would already have marked a large number of clips. It is far from clear that this is the optimal approach to monitoring.

2.2.9.11.4 Manual monitoring function

At least initially, little use appeared to be made of the manual monitoring function; this may have been at least partly attributable to the additional demands that were being made upon senior markers during the early stages of marking. Yet, the function was not always straightforward to use, particularly when supervisors wished to review the marking of questions that were not presently being marked by a marker. In these situations, monitoring required that markers be reallocated to the relevant question (or question group), which could not occur while the marker was on-line.

The Evaluators did not receive evidence of formal (or informal) guidance on precisely how this facility was to be used for evaluating markers.

2.2.9.11.5 Using the monitoring data

There were occasional instances of monitoring statistics that may have led markers to inappropriate conclusions, for example, the computation of mean variance from double marked items. In conventional marking it is usual to compute mean deviations from absolute values (i.e., counting both positive and negative deviations as positive). However, the mean variance statistic aggregated across positive and negative numbers (and lacked any additional absolute component). While this would have been useful for detecting markers who were consistently lenient or harsh, any marker who was wildly inconsistent – but randomly so – would have ended up looking as good as a perfect marker.

As explained by NCS Pearson, such statistics had not been formally defined in advance by QCA and, as such, could not be said to indicate any specific failure on behalf of NCS Pearson. However, examples such as this indicate the kind of concerns that will need to be addressed in future projects.

There were other examples of design features that did not effectively support formative feedback. For example, Netgrade provided no information to adjudicators (within the adjudication function) concerning which markers had awarded the discrepant marks; moreover, to access this information (from another function) could take up to three minutes. The difficulty in accessing this relevant and important information was a source of some frustration to the senior markers. One of the Chief Markers ended up using two computers at once in an attempt to circumvent this problem.

2.2.10 Data processing and storage

There was little in the way of opportunity for evaluating the effectiveness of the data processing and storage systems at NCS Pearson – they were background aspects of the Pilot that were not accessible to direct scrutiny. Indeed, even if they had been more open to scrutiny, the Evaluation team would not have had the technical knowledge to evaluate them. As with any of the more technical aspects of computer hardware, software and programming, it would be advisable to have these separately evaluated by IT consultants before implementing procedures on a larger scale. It will be important to ensure that the systems are effectively designed to achieve precisely what they are intended to achieve, before embarking upon a major scale-up.

However, following the receipt of measurement data from NCS Pearson (analysed in Study 5), a number of question marks were generated. For example, in the revised

databases 01 to 06 (containing item level data from the six papers for all pupils) there were a small number of cases with item marks but without valid IDs (75 cases for one paper). This seems to suggest that there were pupils who became somehow 'lost' in the system. Although this does not imply that they could not necessarily have been 'found' again, it does suggest that the system is not perfect. It would be important to determine why such anomalies occurred and to ensure that procedures were in place for resolving them.

In addition, there were problems with certain of the conventional data (see also 5.3.2.1.2). In the original data feed for databases 01 to 06 there were no missing data for any of the conventional marks. This would, in fact, have been somewhat odd, as markers do tend to make such omissions. Following a request for clarification, NCS Pearson explained that any non-numeric values that had been input (including invalid data and blanks) had been converted to zero. While this was understandable, it was technically inappropriate, as blank cells and invalid entries have a very different meaning from zero marks. Although, of course, these data were merely requested for a research exercise within a Pilot study, this kind of 'data cleaning' confusion could have had major implications if it had occurred in relation to live national curriculum test data.

2.2.11 Results processing and reporting

It was not within the remit of the Evaluation to report upon the results processing and reporting.

2.3 To what extent did the Pilot function as intended?

Accepting a small number of deviations, NCS Pearson managed successfully to implement the agreed procedures for 2001. The most significant shortcomings of the Pilot were:

1. the failure to recruit a sufficient number of expert markers with prior experience of conventional marking;

2. the failure to scan and mark the full complement of scripts;

3. the failure of NCS Pearson to employ the same system (Netgrade) for the Pilot as they would propose to use in a national scale-up (ePEN);

4. the failure to anticipate a breakdown in the on-line supervision process for all marker categories (and the consequent problems for supervisors in completing the processing of DWS sticky notes and Netgrade messages);

5. the failure to anticipate certain of the more frustrating software peculiarities;

6. the failure to accommodate the specific supervisory needs associated with single (rather than double) marking and a consequent failure to provide sufficiently tailored monitoring data;

7. the failure of the absolute-marker reliability ('validity') function;

8. the apparent lack of attention to the possibility of incorrect mark entry (through mis-selection of mark options);

It is important to note that the previous points have been expressed as shortcomings of the Pilot, rather than as shortcomings in the NCS Pearson product, per se. Although it would be fair to say that NCS Pearson were largely culpable in relation to certain of the problems (particularly 3, 5, 7, 8 and, perhaps to a lesser extent, 4) there are others for which the responsibility is to be spread more widely.

For example, NCS Pearson would not accept full responsibility for points 1 or 2, explaining that support in recruiting markers was not sufficiently forthcoming from the Y7PT EMA and that Pilot scripts were not sufficiently forthcoming from markers.

Likewise, it might be argued that the failure of the on-line supervisory process in the UK context was something that could not entirely have been anticipated in advance (and that it was part of the Pilot to evaluate such a process).

Finally, the failure to accommodate the specific supervisory needs associated with single (rather than double) marking is a shortcoming to be shared with the QCA, as this had not received sufficient attention in advance of the project and had not been sufficiently documented in the *Specification*. Yet the identification of this weakness would seem to be one of the more significant outcomes of the Evaluation. Exactly how the training, supervision and adjudication of on-line marking will develop is a major issue that will need to be considered at length by QCA as a matter of priority. It is possible that considerable changes in marking structures will need to be made to adapt to new working practices. Such decisions will need to be considered in advance of future pilots and trials and should be more clearly defined in future project specifications.

The preceding analysis has, inevitably, focused more on weaknesses than upon strengths. Yet, while they have not always been acknowledged explicitly, the strengths of the Pilot are certainly acknowledged implicitly in the recognition that such a novel and ambitious project was largely successful.

Section 3 Markers' impressions of, and attitudes towards, the new marking technologies

3.1 Introduction

Markers' attitudes were investigated in three ways.

1. An initial questionnaire was designed to be given to the expert markers before they started (see Appendix 3.1). Markers who were present on the first day of marking completed this questionnaire very shortly after they began marking. In some cases, markers who started later did not complete the questionnaire until some time after they had started – some were not given the questionnaire until the researcher investigated those not yet received. All fourteen English markers, and eight out of the nine maths markers completed this questionnaire.

2. There were two separate interview periods. The first interviews were completed during the first two days of the marking. The second interview period was in the second week. In the first week, interviews concentrated mainly on those who planned to finish their marking that week. During the three days of the second interview period, the number of expert markers present was fewer than expected, particularly for maths, since there was little work for them to do. However, those who were still present were interviewed, and time was also spent interviewing clerical markers and the Chief and Senior markers. In total, interviews were conducted with twelve English markers, five maths markers, five clerical markers, and all four Chief/Senior markers.

3. The e-markers were sent a second questionnaire after the marking period finished (see Appendix 3.2). This questionnaire was returned by thirteen English markers and seven maths markers.

This report mainly concentrates on the attitudes and responses of the expert markers. The responses of the clerical markers and the Chief/Senior markers are reported separately. A summary of expert markers' responses to the closed questions of questionnaires 1 and 2 are presented in Appendix 3.3 and 3.4, respectively.

3.2 Background

3.2.1 Teaching background

Of the 22 (23) maths and English markers (who completed the first questionnaire), fourteen (fifteen) were practising teachers or headteachers, two were ex-teachers who had become IT trainers, one was a secondary maths advisor, one worked for NCS Pearson but was formerly a secondary English teacher, and four were retired teachers.

Some markers had secondary teaching experience, while others had taught at primary level. Among the nine maths markers, four were or had been secondary teachers (three of maths, and one of biology and ICT); one was an FE maths teacher, and the other four were primary teachers. Seven of the English markers had secondary level experience as English teachers, while the others had a primary teaching background.

3.2.2 Marking experience

One aspect of the markers' background which differed from the original plan was that only seven out of the 23 markers had any previous experience of public marking (despite eight having said so on the first questionnaire). Of these seven, only one English marker and none of the maths markers had experience of marking at key stage 2, although the primary teachers were all familiar with the key stage 2 tests. One of the maths markers had marked key stage 3 maths, one had marked both key stage 3 and GCSE, and one had marked Art GCSE. Two of the English markers had extensive experience of marking key stage 3, GCSE and A-level, while one had marked GCSE only. One, as mentioned above, had marked key stage 2 English.

The markers' backgrounds had some effects on their attitudes which need to be taken into account. The amount of novelty involved for most of them went beyond just the fact of marking on a computer; the majority were also taking part in marking for the first time, and in the case of the secondary teachers, they were encountering a new approach to testing which is very different to that to which they are accustomed. Even for experienced markers, marking intensively at a centre with colleagues and supervisors on hand was unfamiliar, and the pleasant environment of the centre was compared favourably with the facilities available in a school by some markers. In many cases, the markers reported that they enjoyed the marking because it was new and different, or that it was interesting to be involved in a pilot. The challenge is to separate attitudes to the on-line marking from the effects of the other aspects of novelty which all experienced.

3.2.3 IT experience

On the first questionnaire, markers were asked to rate the amount of IT experience they had. Only one marker had never used a computer before, while the experience of others varied. The most frequent uses were for word-processing, email and the internet. The amount of IT experience did not appear to have much effect on the ease of use of the software, as far as the markers' perceptions were concerned – most reported that it was easy to use. Even the one marker who had never used a computer before managed reasonably well with frequent supervision and help, and certainly did not feel inadequate in her use of the system after some initial nervousness.

However, those who had more IT experience were often able to evaluate the system with a more critical eye, and able to make constructive suggestions about improvements. Since they knew more about what is possible, they were able to identify less user-friendly aspects of the system. There were also comments made by some in interviews which suggested that they did not have as good a mastery of the system as they thought, and some of these aspects will be identified in the section below which discusses training needs.

3.2.4 Recruitment/reasons for participation

On questionnaire 1, the e-markers were asked how they were recruited for the marking, and their reasons for taking part. This was explored further with those who were interviewed.

Recruitment fell into 5 main groups (see Table 3.1 below).

Table 3.1 Recruitment for e-marking

Personal contact	8
Applied to AQA for conventional marking	10
Saw NCS Pearson ad in Yorkshire Post	3
Supply Agency	1
Works at NCS Pearson	1

It should be noted that with some of the personal contact group, it was not entirely clear whether initial contact was made through AQA or an advertisement, while others were known by project leaders at NCS Pearson.

Most people gave a variety of reasons for participating and those who were interviewed did not always give the same reasons as they had given on the questionnaire. Table 3.2 summarises these reasons.

Table 3.2 Reasons for taking part in e-marking

Money	8
Useful experience	7
Interested in IT/in being part of a pilot	15
Persuaded	3
Wanted to find out about Y7PT	2

These reasons are discussed further in the next section.

3.2.5 General attitudes

Ideally, it would have been useful to be able to evaluate the markers' pre-existing attitudes to both assessment and the idea of marking on computers. However, the practicalities of the study did not allow for this, since the first data obtained was from the questionnaire which was completed when the markers had been persuaded to take part, had undergone training, and had already started to mark.

The main use of this elicitation of pre-existing attitudes would have been to help judge the extent to which either positive or negative attitudes were influenced by pre-conceived ideas. Presumably, since all agreed to take part, they were not negatively disposed towards the idea – although some had been recruited through personal contacts and reported that they were talked into it, while others had applied for the conventional marking, and agreed to take part in the e-marking when they learnt that they had not been selected for paper marking. In some cases, markers reported motives which were equally served by either type of marking – either the wish to earn extra money or, as was the case with some of the secondary teachers, the desire to find out about the Year 7 progress tests that their pupils will eventually be taking. However, fifteen of the markers specifically

mentioned an interest in IT or in taking part in a pilot project as aspects which attracted them to participate.

The most which can be said about pre-existing attitudes is that at least ten of the markers had applied for conventional marking, so had no particular interest in electronic marking. The personal contact group included three of the markers with previous secondary marking experience. These thirteen can be presumed, therefore, to have a relatively positive attitude towards assessment and marking. Most of those had been told that they had not been accepted as conventional markers before they were approached to take part in the e-marking. For them, therefore, it was not a choice – if they wished to mark, either for the money or the experience, their only option was the e-marking. Only one of this group, however, reported that he initially refused and was eventually persuaded to take part by the Chief Marker for maths; and his reason for refusing was because he did not want to mark at the centre.

Many found the idea of taking part in the Pilot interesting either because they were intrigued by the concept of marking on computer and wanted to see how it worked, or because they liked the idea of being involved with something new. There was a definite perception that this was the future for marking, and a general feeling that it was good not to have to deal with the piles of paper, administration, totalling and packing of scripts which are the lot of the conventional marker.

The conclusion, therefore, is that the majority of markers had a positive attitude towards the Pilot from the beginning. There is no evidence, however, that this attitude was particular to these markers, and would not be shared by others – on the contrary, the fact that more than half had either applied for, or had previously done, conventional marking suggests that this group were not particularly different from the normal markers; and the fact that both new and experienced markers welcomed the convenience of marking without the usual accompanying administrative aspects suggests that the idea may be equally attractive to others.

3.3 Working conditions

3.3.1 Travel

Four of the e-markers were staying in Hellaby Hall hotel as they lived too far away to travel to the centre. Of these, one reported on the first questionnaire that she would have preferred to be going home in the evenings as she had children, but felt that it was better to mark at a centre. In the final questionnaire, however, she expressed a preference for the

flexibility of working at home. Two were happy with the arrangement, while the fourth would have preferred a centre nearer home.

Among those who travelled every day, the shortest journey was 10 minutes, while the longest was an hour and a half by public transport. Those who had longer journeys had generally booked their marking times to avoid rush hour traffic, and most reported that they had planned longer working days both because they did not wish to make the journey too often, and because they would not be able to come in once half term had finished.

Travel and accommodation arrangements undoubtedly affected reactions to some extent, with those who lived further away generally reporting that they would only want to mark at a centre again if it were nearer home. Some worked very long days in an attempt to fit in their 40 hours in as short a time as possible.

3.3.2 Working hours

The final questionnaire asked the markers to report the total number of hours they had worked, and the total number of days. This was a more reliable indication than the NCS Pearson booking sheets, since markers often varied their attendance and the amount of time spent on breaks in a fairly informal manner. Some markers logged off the system when they had no work to do and took a break, while others stayed logged on. The markers' report of their hours is therefore not necessarily the same as the official NCS Pearson record of hours.

Of the twenty markers who returned the final questionnaire, eleven reported that they had worked 40 hours, as planned; two had worked more than this, while seven had worked less. The lowest number of hours was 23, reported by a maths marker. The English markers appeared to have worked the most hours, with two reporting 37 hours, but the rest 40 or more. This was not unexpected since the English team was short of three markers. Note how the high number of hours worked contrasts with the fact that fewer scripts were marked than intended – it seems that the number of 'worked' hours may have included time waiting for marking (at least for some markers).

The lowest number of days was four – reported by two markers who had done 40 hours – and the highest ten. There was a big variation in the maximum number of hours per day, but nine markers had done working days of 10 or more hours – the longest was 12 hours. Of these nine, only three reported that they thought this was too much, although others commented that they did not have any choice. Some felt that the long days were only

possible because they had frequent breaks when the work flow slowed – they would have been very tired if they had worked more intensively.

3.3.3 Working environment

On the final questionnaire, markers were asked to comment on the work stations and the general environment of the centre, and these issues were also discussed in interviews. Generally, they were very positive about both. There were many comments praising the way they had been looked after by NCS Pearson staff, and some said they enjoyed being in a business environment which had much better facilities than those they were used to in schools. Favourable comments were made about the size, temperature, airiness and natural light in the marking room; the comfortable adjustable chairs; the roominess of the workspace; the quality of the screens; the relaxed atmosphere; and the availability of drinking water and snack food facilities.

The main unfavourable comments concerned either noise or physical problems. Seven markers complained that they found it difficult to concentrate at times because of noise from either the scanners or people talking, or both. Ten markers complained of physical problems of some type, including backache, tired eyes, headaches or aching wrists. Although none of these markers reported serious problems, and two of them had pre-existing back conditions, it should be noted that seven of these ten markers were among those mentioned above who had done a small number of very long days. It seems advisable that, if marking were to be done in a similar way again, the health and safety implications of long working hours would need to be considered further, particularly if markers also have long journeys.

3.4 Software/technical issues

On the whole, markers had no problems using the software, and there were very few technical problems reported. Those which occurred were quickly sorted out by NCS Pearson staff. Other issues reported by particular markers were caused by lack of IT experience or the need to learn the software, and were temporary problems which were easily remedied.

However, there were various improvements to the software suggested to solve problems which persisted, or to make it more user-friendly and quicker to use. The issue of delays in the work flow is also included in this section, since this does appear to have been, to some extent, a problem caused by the system as well as by the late arrival of scripts to the centre.

3.4.1 Delays in workflow

The one negative issue which was reported by most markers was delay – the number of occasions when they found themselves with 'No Task' and had to ask to be given more work. Fourteen out of the twenty who completed the final questionnaire complained about spending too much time waiting for work to come through to them, and eight people also complained during interviews, especially during the first week. The majority would have preferred to work much more intensively, although four of those interviewed welcomed the delays as it meant that they had a lot of breaks and did not get so tired. The reasons given for not liking the slow pace were:

- boredom;
- not liking the waste of time, especially since it was half term for some;
- the nuisance of having to find a supervisor to allocate more work;
- loss of momentum and concentration if only a few questions were marked before the work dried up again;
- frustration and loss of motivation.

There were differing opinions on the reasons for the delays. Some felt that the delays came because of late arrival of scripts to the centre; others had heard that it was because markers were faster than expected; while others believed that part of the problem was that the process of re-allocating questions to markers had to be done by NCS Pearson personnel or a Chief/Senior marker, that this process was not simple, and they were not able to keep up with the number of changes necessary.

The interviews with the Chief and Senior markers led to the conclusion that the problem was caused by a combination of: late arrival of scripts (so that scanning did not start early enough); the slowness of the scanning process; the fast pace of the markers (especially since the pupils taking the test were low achievers); and the complicated process needed to re-allocate markers. One compared this process to painting the Forth Bridge, and it was an obvious frustration for all of them.

Late arrival would presumably be less of a problem with live marking where scripts would come straight to a scanning centre. Also, the marking would probably not be so fast in a test which covered a wider ability range. The question re-allocation issue –

which is discussed further when the Chief/Senior markers' interviews are described – is one which appears to need a software solution.

3.4.2 Scanning

The markers were generally positive about the clarity of the scanned image – only five markers reported problems. In one case, this was a problem with a maths question where lines drawn on a diagram could not be seen distinctly. There were also a few problems caused by annotations made by the paper markers.

The biggest issue was the scan area; either during the interviews or in response to questionnaire 2, thirteen markers mentioned the problem of answers which had been missed because the child had written outside the scanned area. Although they had been told to mark only what they could see on the screen, many markers were not happy doing this (even though they realised they were not really awarding the pupils' final marks) as they still felt they were being unfair (see also 6.4.1). Some followed the instruction to mark it anyway; some sent a message to the Chief/Senior marker; while some flagged the questions.

This is a problem which would obviously need a solution, since markers felt so uncomfortable about giving a mark when they knew they could not see the whole answer. If this had been live marking, they would have been very unhappy about it – and probably unwilling to impose a strict rule which penalised children, even if they knew that children had been warned not to do this. One thing which united all markers was a desire to be fair to the children who took the tests, and anything which is perceived as unfair would probably lead to resistance – as in this case, where most did not want to take the action they had been told to take.

3.4.3 Mark scheme

Instructions given to the markers, concerning the use of the on-line mark scheme, varied. The maths markers were told at the training that they should use the on-line mark scheme exclusively and should not bring the paper mark scheme to the marking, although there was some subsequent confusion about whether they were actually allowed to bring the paper mark scheme. Most ignored the instruction and brought it with them anyway – one reported at interview that she had it outside in the car in case of emergency! The English markers were told to bring the paper mark scheme. Not surprisingly, use of the on-line mark scheme was greater among the maths markers than the English markers. On the final questionnaire, only one maths marker reported using the paper mark scheme more

than the on-line mark scheme, while only one English marker used the on-line version most of the time. Another English marker reported equal use of the two versions.

Among the reasons mentioned at interview for not using the on-line mark scheme – reservations which were reported by some of those using it as well as by those who were not – were:

- The annotations made during marker training. One English marker referred to the personalised version of the mark scheme as 'an ally', while another said they were told at training that the annotated version became 'a bible'. The maths marker who used the paper mark scheme also mentioned the annotations made as the main reason for doing so.

- The limited screen space and the difficulty of seeing the mark scheme and the pupil response on the same screen. Although it was possible to re-size the mark scheme window, it was fiddly to do so while still being able to see enough of the mark scheme at the same time as the response. Some of those with limited IT experience either did not realise this was possible, or did not work out how to do it, even though it was mentioned during the training.

- The fact that the mark scheme window disappeared as soon as the response window was clicked on. This was particularly a problem with the English writing question, where scrolling was necessary – every time the marker scrolled, the mark scheme disappeared, and keystrokes were necessary to bring it back. Also, the mark screen disappeared each time an item was marked, and had to be brought back for the next one.

- With the writing question, the need to look between the pupil's writing, the mark bands and the examples several times before a decision could be made about a mark. This was easy to do with all visible at once – but very awkward when it involved a lot of mouse clicks and key strokes.

The present version of the mark scheme appears to be most useful with questions which have short answers, demand a limited amount of judgment from the marker, and are easy to mark. In these cases, the marker becomes used to the mark scheme and refers to it less and less. With more complex questions, either the marker will spend a lot of time recalling the mark scheme, or will fail to do so often enough and may thus mark less reliably. The problem is alleviated to some extent because markers mark a lot of the same item at once, so are more likely to remember the mark scheme. However, with items

which need continued reference to band scales or examples, the present design is unwieldy. The marking of the English writing, in particular, would need considerable thought to develop a design which was acceptable to markers as an on-line version.

The part of the mark scheme viewed most favourably by the English markers was that for the handwriting, where the ability to compare the examples and the response on-screen, one above the other, was seen as beneficial – but only if the problem of the disappearing window could be solved. At the very least, the facility to bring it back with a single mouse click rather than alt-tab would be useful.

In addition, the issue of markers' annotations needs to be overcome for the on-line mark scheme to become totally usable. A way of adding to the mark scheme and thus developing an annotated version – and of doing this in conjunction with training, which implies on-line training – would be needed. An alternative would be to continue to use the paper mark scheme, or a combination of the paper and on-line versions.

3.4.4 The message system

The message system proved inconvenient as Chief/Senior markers could only receive a message when logged on to the question which the marker was marking at the time of sending the message. The problem was compounded by the fact that, particularly in the beginning, the Chief/Senior markers were busy either re-allocating markers to questions or working with markers at their stations, helping with either use of the software or with marking issues. This meant that in the beginning, markers were sending messages but not getting answers. Eventually, markers stopped using the message system – and relied on face-to-face support – unless they had concerns that did not need an immediate response.

3.4.5 Looking for work

When markers came to the end of a batch of questions – or when they wanted to switch task – they initially went back to the original log-in screen to do this. This meant that they had to exit, log out, re-enter their password and log in again onto a new question. If they happened to select a question where there were no unmarked items waiting, they had to redo this process. This was mentioned by many as being a waste of time and, for some, a cause of frustration. Some accidentally discovered another way of doing this which was not mentioned in the user guide, using a button on the title bar of the main menu labelled 'Select function'. Use of this method then spread by word of mouth to others, but some never discovered it and continued to complain about the laborious method.

This is to some extent a design issue – in particular, giving the button a more obvious name and including it in the User Guide – and to some extent a training issue. Since the NCS Pearson staff did not appear to know this method, it will presumably be sorted out as they become more familiar with the software. Alternatively, some markers suggested that the process would be easier if 'back' buttons were used, similar to those on a web browser.

Another change which a number of markers suggested was that it would be useful to be able to quickly see if questions were waiting to be marked, and also how many there were to mark. Some markers like to pace themselves by knowing how many more they have to do – some saying that, if they knew they only had a few to go, they would finish before taking a break; while if there were a lot, they would take a break then continue. It seems that markers would value being able to have this flexibility; and if marking at home, this facility would help them manage their time.

3.4.6 Back-scoring

When a marker wanted to go back to a question previously marked, the back-scoring system was used. This enabled the marker to see the last 10 items marked (this was changed during the marking process so that they could see 20 rather than 10).

The markers had no problems accessing the system, but a few modifications were suggested:

- It would be useful to have a numbering system on the back-marking selection screen which identified the items more clearly. The last one marked was always at the bottom, and became number 10 or number 20 – on the next access, it would be 9 or 19 – so finding an item a few back was often a matter of trial and error.

- Some markers mentioned that they would have liked to have had a way of 'setting aside' items with queries which they had asked about in a message to the Chief/Senior markers. The present system means that they had to commit themselves to a mark even when they were unsure – and if an answer to their query came later, they could not go back to the item. Although they believed that such items would probably be picked up again later as part of the checking process, they still worried that this might make their marking look unreliable. This was one of the few issues compared unfavourably with paper marking by those who were familiar with the paper marking system, where they could use post-it notes or set aside such questions until they had

been able to phone or e-mail their team leader. The markers' pride in doing a good job – and their wish to take responsibility for their marking – was at issue here.

3.4.7 Terminology

There was a general feeling among markers that the terminology used for the various software functions and messages needs to be made clearer. This feeling was expressed partly by the jokes being passed between markers as they compared strange terminology they had encountered – sometimes accompanied by jokes about translation from Chinese, since they knew that this software had originally been used in China. Specific terminology mentioned during interviews or on questionnaire 2 were the terms 'scoring client', 'open error', 'swapped from practice to live', and the term 'select function'. With the exception of 'select function', none of the terminology appeared to have caused any problems; but it undoubtedly increased the learning load, confused those who were less IT-literate, and would have caused more problems if there had not been immediate technical support available. If the software were to be used more widely, and particularly if markers used it at home without immediate help available, it would seem wise to make the terminology used more transparent and user-friendly.

3.4.8 Other issues

- The zoom option was mentioned favourably by several markers. English markers found it useful when marking handwriting, and one maths marker had found it useful to check numbers when handwriting made them ambiguous.

- One marker found a use for the 'rotation' option when a script had been scanned upside down.

- Six markers had problems with selecting the mark with the mouse, particularly when marking questions quickly. This seemed to be a common reason for using the back-scoring function. Selecting a mark needed a mouse click in a small circle, and those who had less computer experience, in particular, did not always click positively enough, the result being that the item was marked as 'no response' (the default). They often realised this after (or as) they clicked the 'Commit' button, when they noticed that the intended selection had not turned red. They then had to go to the Back Scoring screen to correct their mark.

3.5 Changes in attitudes

Six items from questionnaire 1 were repeated on questionnaire 2, to estimate changes in general attitudes to computer-based marking. Table 3.3 summarises these responses. Since numbers were so low, responses have been combined to show positive, negative and neutral attitudes to each item. The questionnaire 1 responses of the two markers who did not return questionnaire 2 were checked. Their responses were all in the positive range on questionnaire 1, so their absence does not account for any of the improvement in positive attitudes. One marker did not fully complete these items on either questionnaire 1 or questionnaire 2, which accounts for totals which are less than the total number of returned questionnaires.

The general trend was towards a more positive attitude on all but the last item. The improvement was particularly marked in the case of speed of marking and the marking of individual questions rather than whole papers; also more had enjoyed the experience than had expected to. Reading from a screen, however, was not as positively perceived by the end. As already mentioned, some complained of eye problems or headaches, and many had long working days. This tiredness factor may underlie these responses.

Table 3.3 Comparison of questionnaire responses

	Questionnaire 1 (n = 22)			Questionnaire 2 (n = 20)		
	- ve	neutral	+ ve	- ve	neutral	+ ve
Computer-based marking is quicker	2	2	17	-	1	19
Preference for computer-based marking	1	2	18	-	3	16
Marking questions separately v. marking whole paper	2	9	10	-	4	15
Enjoyment of marking experience	-	9	13	-	3	17
Interest in marking experience	-	5	17	-	1	19
Reading from a screen v. reading from paper	6	4	12	4	7	9

A further confirmation of improvement in admittedly initially positive attitudes comes from the item on questionnaire 2 which asked the markers to rate the marking experience as better or worse than they expected. Eleven markers reported that the experience was better than they expected, while the other nine rated it as neither better nor worse. Reasons given in written comments were:

- People factors, such as the company of fellow markers and friendly and helpful NCS Pearson staff. This was the most common reason given for positive comments.
- The marking was easier than expected, with fewer problems using the computer than feared, or less stress and tiredness.
- The experience was interesting and stimulating.
- It was good to be involved in something new.

The first point would not, of course, be an influence if computer-based marking was eventually done at home. There was a noticeable social aspect to the marking, fuelled to some extent by the fact that the breaks in the work-flow, particularly in the beginning, gave markers plenty of opportunity to get to know each other. In addition, the markers responded well to the comfort of their surroundings and the general attention which was paid to the whole Pilot process by all involved. They had a tolerance of problems which came from the view, expressed by many, that any pilot would have teething problems, and a general goodwill to co-operate to make things go well. The number of visitors at various times, and the presence of NFER researchers who were interested in their feedback, no doubt added to the impression of being involved in something new and important.

Perceptions of the fairness of the payment were also reported on both questionnaires. On questionnaire 1, ten markers gave positive responses, while ten were neutral. One did not respond, while the remaining marker was an NCS Pearson employee so the question was not applicable. On the second questionnaire, one was negative, ten were neutral and eight were positive. Examination of the pattern of responses does not suggest much change in reactions to the payment – it was initially considered reasonable or good by most, and they continued to think this. The only negative remarks about payment were made by a few who commented that it would be fairer to pay by the number of items marked rather than by the hour, since this would reward more experienced markers. However, since markers were not specifically asked this question, this cannot be said to be a general view.

3.6 Perceived advantages and disadvantages

On both questionnaires, markers were asked to list the advantages and disadvantages of marking on computer, both for markers and for the education system more generally.

3.6.1 Potential advantages for markers

Responses fell into five main areas.

3.6.1.1 Monitoring/supervision/mutual support

Responses in this area mainly concerned issues specific to marking at a centre, although mention of on-line monitoring and feedback has also been included in this category. There were four mentions of advantages in this area on questionnaire 1, and seven on questionnaire 2. Two of those who mentioned an advantage in this area on the first questionnaire were also among those who mentioned it on the second. The increase in responses on the second questionnaire was mainly accounted for by those who mentioned mutual support from colleagues, so the influence of their enjoyment of the social aspects of the marking is again evident here.

3.6.1.2 Less administration, fewer piles of papers, etc.

This was the area most frequently identified as an advantage – mentioned by thirteen markers on the first questionnaire, and fourteen on the second. This is in spite of the fact that the majority had not marked before; they nevertheless understood the present marking arrangements well enough to appreciate the benefits of a paperless system.

3.6.1.3 Benefits of marking one question at a time

Nine people on questionnaire 1, and nine on questionnaire 2 mentioned advantages in this area – although they were not the same nine. Overall, fifteen markers mentioned advantages which were mainly concerned with the ease of building up expertise in particular questions, or the increased accuracy of marking.

3.6.1.4 Speed

Eleven people on questionnaire 1, and nine people on questionnaire 2, mentioned faster marking as an advantage.

3.6.1.5 More accurate/reliable/etc.

Advantages in this area were sometimes linked with the first category – the ease of monitoring – and sometimes with the benefits of marking individual questions. Eight mentioned advantages in accuracy and reliability on the first questionnaire, and seven on the second. In all, thirteen gave this as an advantage in one or both questionnaires.

3.6.2 Potential disadvantages for markers

There were fewer mentions of disadvantages than advantages. The disadvantages mentioned fell into six main categories.

3.6.2.1 Technical/software problems

Possible problems were mentioned by three on questionnaire 1 and six on questionnaire 2. Comments on questionnaire 2 were generally more specific than those on questionnaire 1, mainly mentioning software issues or delays which have already been described.

3.6.2.2 'Big brother' aspects, erosion of professionalism, loss of 'the feel for the child'

Concerns in this area were mentioned by four markers on questionnaire 1, but only one (a different person) on questionnaire 2. To some extent, this was part of the process of getting used to marking key stage 2 tests. During interviews, the only experienced key stage 2 marker mentioned that she had noticed some of her fellow markers having problems adjusting to the objectivity required, and the need to keep to the answers in the mark scheme. She herself had gone through the same adjustment when she first started marking.

3.6.2.3 Boredom and tiredness, when marking one question

Four people mentioned this as a possible problem on questionnaire 1, and four different people mentioned it on questionnaire 2.

3.6.2.4 Scrolling/reading on-screen

This was mentioned by three on the first questionnaire, and a different person on the second. Not surprisingly, all were English markers.

3.6.2.5 Physical problems

Eye or wrist strain were mentioned by two people on questionnaire 1 as potential problems – neither of these two mentioned this issue on questionnaire 2. However, five people mentioned either specific physical problems or physical tiredness on the second questionnaire.

3.6.3 Potential advantages for the education system

The number of advantages mentioned on the second questionnaire was greater than that mentioned on the first, which may suggest that the markers' experience of the process increased their positive impression of the benefits.

3.6.3.1 Speed of marking/faster results to schools

This was the most commonly mentioned advantage, given by twelve people on questionnaire 1, and fourteen on questionnaire 2.

3.6.3.2 Reduced administration and movement of papers

This was mentioned by five markers on questionnaire 1, and nine on questionnaire 2.

3.6.3.3 Reduced cost

Mentioned by four people on questionnaire 1, and six different people on questionnaire 2. Most of those who mentioned this were not sure, however; they thought it might be cheaper, but some added comments that they did not really have enough information to judge this.

3.6.3.4 Easy provision of national or school data

Seven people on questionnaire 1 and seven on questionnaire 2 mentioned this – a total of nine people on one or both questionnaires. In particular, the benefits for a school in having data which could improve teaching were mentioned.

3.6.3.5 Marking fairer/more reliable and accurate

Ten people on the first questionnaire and six on the second mentioned this – mainly connected with the ease of monitoring and double marking. This is the only issue where responses were less common on the first questionnaire; of those who mentioned it on the first, only three mentioned it on the second.

3.6.4 Potential disadvantages for the education system

There were more disadvantages mentioned on the second questionnaire than the first. This increase was mainly accounted for by technical concerns that arose during the trial.

3.6.4.1 Hardware problems/system crashes/viruses/power cuts/etc.

This was mentioned by five on questionnaire 1 and six on questionnaire 2 – nine people in all on one or both questionnaires.

3.6.4.2 Scanning problems

Mentioned by two on questionnaire 1, and six different people on questionnaire 2.

3.6.4.3 Insufficient IT skills of markers

Mentioned by two on questionnaire 1, and two different people on questionnaire 2.

3.6.4.4 High set-up costs

Interestingly, this was a disadvantage which occurred to markers only during the marking, since none mentioned it on the first questionnaire, whereas eight mentioned it on the second.

3.6.4.5 Other aspects

Another aspect which was mentioned on questionnaire 2 only was the 'de-humanising' aspects of computer marking, mentioned by two people. Also, one person referred to the problems associated with the NQT tests, and another referred to the need to depend on 'computer organisations' – possibly a result of there having been some discussion at the marking centre of the NCS Pearson involvement in the NQT tests.

3.7 Marker culture

This section describes more general issues concerned with the experience of the e-markers. However, the cautions given earlier should be repeated here: for the majority, this was their first experience of marking of any type and only one had marked key stage 2 tests before. They are, therefore, reacting to many aspects of the newness of the situation.

3.7.1 Centre v home

On the first questionnaire, sixteen out of the twenty-two respondents said they would prefer to mark at home. On the second questionnaire, a 'maybe' option was added. Twelve preferred home, four preferred a centre, and four responded that 'maybe' they would prefer home.

The most common reason for preferring to mark at home was that this would mean being able to be flexible about working hours and working at one's own pace. Twelve gave this as a reason. This issue was mentioned by many during interviews, and was often prompted by the perception of their time being wasted when delays meant they had no work.

Other reasons either given on the questionnaire or mentioned during interviews were the quietness, comfort or facilities of home, and the travel time needed when marking at a centre. The fact that it was half-term was mentioned by some, particularly by those who had children. They would have liked the flexibility to be able to go out when the sun was shining, take a break when they felt like it, and work late at night or early in the morning if they wished.

However, this does not indicate a preference for conventional marking – there was generally great support for the idea of marking at home on computer, with the associated advantages of lack of paper and reduced clerical work. The only hesitation was caused, unsurprisingly, by technical concerns.

The minority who preferred to mark at a centre mainly mentioned the lack of distractions which made it easier to concentrate, and the discipline of working set hours. Some mentioned interruptions from their family as a distracting influence at home. The immediate availability of help from supervisors or colleagues was also mentioned as an advantage of being at a centre – which was perhaps partly exaggerated due to the fact that they were inexperienced markers.

So, although the e-markers enjoyed the experience of marking at a centre, the majority would nevertheless prefer to mark at home. Some mentioned that, if the marking were to be done at a centre, they would prefer it to be at one nearer their home. The option which was taken during the Pilot – of paying for hotel rooms or paying travel expenses for what were in some cases long journeys – would, in any case, presumably not be economically viable on a large scale.

It should also be noted that, according to the Chief Marker for maths (who, as an NCS Pearson employee, had also been involved in recruitment for English), the main reason for people not wanting to take part in the marking was the travel involved and the time that would have to be spent at the centre.

3.7.2 Secondary school teachers

There was a particular view – which was generally expressed as 'losing the feel for the child' – which seemed to be shared only by some of the secondary school teachers involved in the marking. It was not always easy to separate how much of this view came from a reaction to differences in the approach to assessment at key stage 2, and how much was a reaction to e-marking. This was undoubtedly a reaction to some extent against the approach to assessment. This arose from the increased objectivity required to keep to the mark scheme, which some felt was an erosion of their professional judgement. The fact that they were marking individual questions rather than whole papers increased their concern with losing the feel for the child.

However, as mentioned earlier, the one experienced key stage 2 marker reported at interview that this was a common attitude among new markers – she remembered having to get over it herself and learn to mark objectively. It may well be, then, that this is a training issue, and that it may in any case decrease as secondary school teachers become more accustomed to the Year 7 Progress tests.

Another point raised by those who had marked key stage 3 or higher level exams was that, although they felt that e-marking was suitable for key stage 2 or the Y2PT, they were not so sure it would be suitable for questions with longer answers which would require a lot of scrolling to read.

3.8 Clerical markers

Four out of the ten maths clerical markers and one of the two English clerical markers were interviewed. Four of these were students and one was a retired maths teacher. They were recruited for the marking through the Adecco agency, which gave them a keyboard test, and had a training session at NCS Pearson which covered both the mark scheme for the questions they were to mark, and the system and software. They all reported that they had found the training useful, and had no problems with the software.

The issues which arose from these interviews were:

- They all felt that they had benefited from plenty of help and supervision.

- They found the working environment comfortable. The only complaints were that the air-conditioning was occasionally too cold and, in two cases, that their eyes became very tired. One attributed this to the bright screen colour.

- The maths markers had spent a lot of time waiting for work. They perceived this as arising from the 'logjam' which occurred in the system when they sent queries to supervisors. They were under the impression that scripts with queries remained in the system until they were dealt with, and this meant that new clips could not come through.

- The maths markers had spent some time doing data entry when there were no clerical questions to be marked. Only one objected to having to do this, although the others said they found it boring.

- The most common reason for sending queries to supervisors was scanning quality – either too faint to read, or answers which were written outside the scanning area.

- They would have liked a way of going back to items previously marked when they realised they had made errors. They were able to do this where they were marking several items on the same script, but could not go back to a previous script.

- Three out of the five reported that it could be difficult to maintain concentration – 'You get to the stage where you're in trance', as one said.

- The use of '\' to mark a blank was mentioned. They felt it would be better to use '/' since this was on the right hand side of the keyboard, and they were using the number pad with their right hand.

- One maths marker mentioned that it was not possible to enter a minus when doing data entry – the software would not accept it. Such cases had to be flagged for the supervisor.

- The English marker reported that some questions were difficult to mark without having the mark scheme visible at the same time as the question. In particular, a matching question which required checking lines drawn by the child needed frequent checking of the mark scheme. She had eventually been given a copy of the paper mark scheme to help with this.

Generally, the clerical markers seemed to have a positive attitude towards their work, and were happy with the payment they were receiving (especially the students, who were paid more than they would be for other types of holiday work).

3.9 Supervisors

There was a Chief Marker and a Senior Marker for each subject. The Chief Marker for English was an experienced key stage 2 Senior Marker, who had recruited as Senior Marker one of her Team Leaders. The maths Chief Marker was also the NCS Pearson Examinations Manager and had extensive experience of maths marking at key stage 3 and GCSE. The maths Senior Marker also had key stage 3 marking experience, and had worked as a QCA consultant on key stage 2, administering and marking pre-tests. All four of them were interviewed and were also spoken to informally at different stages throughout the Pilot.

The first thing to note is that all four were very committed to making the Pilot successful; the good team spirit and pleasant environment commented on by so many markers was in large part due to these four supervisors. They worked long hours and spent a lot of time supporting and helping the markers. This was particularly necessary because so many of them were inexperienced.

What they were not able to do properly, however, was the role that they were supposed to have as e-supervisors – which was the on-line checking and monitoring of the markers. In neither subject did this happen until the second week, by which time many of the markers had finished.

The impression was given at both the maths and the English training that the supervisors would be on-line during the e-marking, and that their main role would be to check, monitor and to deal with any queries which came to them through the message system. This broke down on the first day, partly because work-stations were not always available, and partly because they became involved in re-allocation of questions. It eventually emerged that the messaging system could not be used by markers to get help quickly, since the supervisors only got messages when logged on to the specific question for which a query had been raised. One supervisor reported that they were eventually told 'by NCS Pearson' not to bother with the messaging system, as it was unnecessary while they were in the room.

The supervisors did not all feel they were totally well-prepared for their role. One had not understood the full nature of the project, including the difference between the various

types of markers, until the end of the first week; one who had limited IT experience had a lot of problems with the software; another would have liked more training on the monitoring and checking system, finding this very complicated to use. They felt uncomfortable when they were unable to answer markers' queries about the reasons for delays, and there was an anxiety which came from a feeling of being de-skilled and having to constantly ask for help themselves before they could help the markers.

All four of them had spent the first few days of marking in what one described as 'crisis management'. They had, in a sense, taken responsibility for trying to keep the markers happy while at the same time trying to keep the work flowing in any way they could – even if this meant abandoning the supervision system as it had been designed.

As already mentioned, a lot of their time was taken up re-allocating questions to markers. They mentioned the length of time it took to do this, and the fact that this had not originally been part of their role. The plan was that NCS Pearson technicians would re-allocate all markers to different questions every two hours. However, questions were not available for marking as soon as they had been scanned, scanning had started late, and more had been rejected by the scanning process than had been expected. These delays were compounded by the fast marking (which may well have been somewhat attributable to the low ability of the pupils). As a result, the re-allocation system did not operate as planned, since questions had to be re-allocated constantly as markers ran out of work. During the first few days of marking, it seems that the supervisors' time was divided between re-allocating questions and answering markers' queries face-to-face. All four agreed that there was a need for an automated system to allocate items to markers.

In the second week – which is when they were interviewed – things had settled down because there were few markers present. The supervisors were then able to turn to checking marking and dealing with queries, as well as to adjudication of clerical marking. They were all very appreciative of the advantages of being able to do this type of supervision, but regretted that they had not been able to do it in the beginning when it could have given useful feedback to markers. The system was, however, regarded as very complex at present, and there was a general feeling that it needed to be more user-friendly. In particular, the need to constantly log on and off was time-consuming – the English markers, for example, needed to do this 26 times to check all questions. Separate log-ons were then needed to adjudicate clerical markers' work.

The supervisors believed that the monitoring results had shown a reasonable amount of accuracy among the expert markers, but a much lower rate among the clerical markers. They perceived a lot of errors with the marking of spelling, some described as 'blatant

errors'. The English Chief Marker felt that spelling should be marked by the expert markers. There were also a lot of errors in the maths clerical marking, some due to confusion between marking zero or blank, but others due to 'incorrect interpretations', or possibly going too fast.

There were some reservations about the idea of supervising markers working at home in this way. These reservations all appeared to have their origins in the fact that they had not, in fact, been on-line supervisors during the Pilot – they had been doing face-to-face supervision of markers who were working on computers. Firstly, it was pointed out that a more efficient message system would be needed, which operated more like e-mail, without the requirement for a separate log-in to each question, and with the facility both to know when a message was waiting and to reply to messages easily. Secondly, they felt that software modifications and further piloting would be needed before the system could be scaled-up – especially for use by markers in their own homes. Thirdly, they had worries about the loss of the social aspect of marking, especially if training was also done on-line.

There were a few other issues mentioned, mainly to do with technical or software matters:

- Terminology was obscure and needs to be made more familiar and everyday.
- Scanning problems were mentioned, similar to those mentioned by the expert markers.
- The question numbers on the paper and the ID numbers in the system were not the same, meaning that a list had to be consulted when questions were being re-allocated.
- The page numbers on the scanned paper were not the same as those on the printed paper, and pages could not be scrolled through. This meant that time was sometimes wasted when looking for a question.
- The re-allocation process was lengthy and involved several steps – and could only be done when a marker was logged off from the system.
- There appeared to be doubts about what happened to clips which markers had flagged. They had been told by a manager at NCS Pearson that they 'went back into the system', but nobody seemed to know how to find them.
- Data entry queries and clerical marking queries came into the same system, and there was sometimes confusion about which they were. (This, of course, would not be a problem with live marking where there would not be data entry.)

- The mark scheme needs some re-design. In particular, the need for personal annotation; the need (which would continue for English markers) to have access to the reading booklet; and a suggestion that the 'Do not accept ..' part of mark schemes should be at the top, not the bottom.

3.10 Conclusions

3.10.1 Markers' willingness to participate

The markers included in this study had a positive attitude towards the concept of marking on computer, and could see great benefits for markers and for the educational system. However, they were less enthusiastic about marking at a centre, despite the fact that they enjoyed this experience. To some extent, their enjoyment came from the novelty of being involved in something new, the comfortable conditions in which they were working, and the social aspects of meeting new colleagues. However, the flexibility of working from home still attracted them. This was partly affected by the delays they had experienced and the distance between their homes and the centre. A centre closer to their homes would probably make the prospect more attractive.

3.10.2 Technical issues

Various issues have been mentioned throughout this report concerning the usability of the marking system. Although the markers found the Netgrade interface easy to use, there were nevertheless various aspects which they and the supervisors would like to see improved. In particular, the messaging system, question allocation, the monitoring system and the comprehensibility of the terms used need attention. Markers would also like to be able to see how many questions they have in a batch still to be marked, and would like to be able to see more easily when they have questions to be marked.

There are also scanning issues which need to be resolved – the speed with which questions come through to markers once scanned, making sure that the whole of a pupil's answer is scanned, and a way of dealing with images too faint to read.

3.10.3 Training

Some of the problems which markers had with the software could have been easily avoided with more time spent on hands-on training. Markers commented that not many practice questions were available on the training day. This was probably because few had been received for scanning at that point, so would not necessarily be the case in future marking exercises.

Supervisors appeared to need more preparation for their role, particularly if the monitoring system remains as complex as it is at present.

3.10.4 Question allocation

Allocation of questions to markers did not go as intended, partly because of the speed of marking, and partly because of the slowness of scanning and of the process of allocating questions. It seems rather unwieldy at present, and was the main source of frustration for markers and supervisors. It either needs a software solution, a much quicker method of manual allocation, or a means by which markers can access questions directly.

3.10.5 Mark schemes

Lengthy mark schemes, such as those for parts of the English papers, proved not to be very usable with this system. For shorter items, markers would prefer a mark scheme which remained visible on the screen. For all items, a way for markers to add their individual notes would make the mark schemes more usable.

3.10.6 Marker fatigue

There were signs that the very long days worked by some markers caused excessive tiredness, headaches, sore eyes and other physical problems. If marking were completed at a centre near home, this might pose less of a problem since the long days were a practical necessity for some – either they lived too far away to come in for short periods, or they were staying in a hotel. It would seem wise to restrict the maximum number of marking hours, both for health and safety reasons and because fatigue could adversely affect the accuracy of marking.

3.10.7 Clerical marking

Supervisors believed that an undue number of errors had arisen from the non-expert markers, for questions that should have been very simple to mark. In future projects, they would need to be reassured that marker training was sufficiently rigorous. In addition, in light of concerns over spelling, it will be important to ensure that senior markers fully support the categorisation of all questions.

Section 4 Stakeholders' impressions of, and attitudes towards, the new marking technologies

Section 4 summarises findings that arose from interview studies with six key stakeholders. The interviews explored their impressions of, and attitudes towards, the new marking technologies and the implementation of such technologies in the Pilot. The six participants were as follows:

- A DfES Senior User on the Joint Project Board, with responsibility for the National Data Collection. He was a technical user of the national curriculum assessment (NCA) data, using it in many ways (schools' results, Autumn Package, etc.). At times, he also provided a broader DfES viewpoint.

- The QCA Senior User who was, effectively, the sponsor of the project (which he saw as central to re-vamping NCA over the next few years). He was not involved in the project in detail but was generally aware of the issues.

- A representative of the National Data Collection Agency (NDCA), who was responsible for managing the process for Edexcel. He had a limited role in the Pilot.

- A representative of the External Marking Agency (AQA), who was involved in drawing up specifications and making arrangements for scripts to go to NCS Pearson but who has not seen detailed processes. AQA was to provide markers and scripts for NCS Pearson.

- Two interviews were undertaken with NCS Pearson representatives: the Managing Director (who was also a member of QCA's Project Board) and the NCS Pearson Project Manager. Both had a full grasp of the detail of the project.

Hence these six divide into three pairs, the commissioners (QCA, DfES), the current operators (AQA and Edexcel) and the aspiring suppliers (NCS Pearson).

Two sets of interview studies were conducted with these participants. The first, conducted towards the end of June, explored aspirations for the project and initial views on the functioning of the Pilot. The second set conducted during the beginning of September, was intended to be more reflective and to explore opinions in the light of a fuller understanding of how the Pilot had functioned. In fact, little more emerged from the second interviews, as most participants had not had an active involvement with the project over the intervening period. Some had received management reports, but with little detail.

4.1 Expectations of the Pilot

All contributors were agreed that there is a substantial need for new technology to be introduced into national curriculum assessment. This is either because of an aspiration for greater efficiency in reducing times taken, while increasing accuracy and reliability, or to overcome problems with the current process – the lack of markers, costs of training or large number of requests for reviews.

DfES made the point that although strong advocates of such systems, NCS Pearson did not provide sound evidence of consistency of marking (etc.) from their US experience. No research or evaluation evidence of the proposed benefits has been provided.

There were a range of degrees of desire for change, from NCS Pearson seeking a definite commitment (to a programme of introducing e-marking) through the cautious approach of the commissioners, to the views of the current operators (that much could be done with safer, already tried technology – OMR capture of markers' data, clerical additions, etc. – which they utilise in other examination work).

The Business Case had identified scanning as the centre of the process and an area from which all else flowed. There was, therefore, a need to demonstrate that the expectations of this were sound.

Underlying many comments and attitudes, and referred to spontaneously by all, was the 1998 score sheets "disaster". For the DfES there is a wish to avoid public mistakes and this transmits itself to QCA and their contractors. There is a real tension between this caution and the desire to introduce new technology. NCS Pearson are perhaps the most realistic in asserting that there will be a bumpy ride but that it must be expected and managed, if the benefits are to be realised. This contrasts with the current operators' views (from a background of delivering traditional examinations) that an implementation must be perfect from the start to guard standards and the life-chances of individuals. For the management of the future developments it would be important to agree and decide on level of risk tolerated at each stage and share this with those involved. This issue is explored further in Section 4.4.

4.2 Operation of the Pilot

4.2.1 Obtaining scripts

The throughput of the Pilot was much less than planned. The number of scripts did not emerge on time and not early enough to keep markers busy over the main marking weeks

(half term). This was explained by AQA as being due to confusion in schools over the Year 7 progress tests and the optional Year 7 tests. The late start to the process and mixed messages from QCA and DfES (particularly concerning the participating schools) contributed. NCS Pearson, on the other hand, felt that AQA had not been effective in ensuring the supply of scripts and markers. A consequence was that the volume of marking was not stretched and the capacity of the system remains untested. This is a major limitation of this pilot. In further pilots a change to the flow of papers (if a parallel operation is used), where scanning is undertaken before conventional marking, would reduce the problem of work flow to electronic markers.

Participants felt that the technical aspects (such as the personalisation, despatch to schools and tracking the scripts back) appeared to work well. In later interviews NCS Pearson acknowledged that there had been problems with the printing, but asserted that these have now been resolved.

4.2.2 Adapting scripts for scoring

Very little adaptation of the questions was permitted by QCA. This was a source of some frustration at NCS Pearson, who believed that slight adaptations in answer format (in both Maths and English tests) would allow automatic marking of some questions, which are currently disguised multiple-choice. Others, particularly QCA, did not wish to see any change to the tests. This issue was important in relation to the case for saving costs. Large cost savings can only be accomplished with the removal of human markers, implying changes in format to many questions. However, for reasons of validity, the QCA board (and perhaps the education system) would not tolerate changes which appeared to be making the tests exclusively multiple-choice. This indicates the need for a debate on the extent to which such questions (those markable by machine following scanning) are to be permissible and, if they are to be used, on their nature. The costs and benefits need to be agreed so that all are clear on the parameters to which they are working, from test development onwards. An issue which has thus far had very little consideration, is the linkage of the marking process to the earlier stages of test development, printing, distribution etc.

The adaptation of other aspects, such as the page markers etc., worked well, apart from in a relatively small percentage of cases where children doodled or shaded in these areas, confusing the scanning.

4.2.3 Scanning process

Little information emerged about this from the interviews. The current contractors, AQA and Edexcel, were convinced that data will be lost, bundles of papers dropped and become unscannable or that other disasters will happen, threatening the validity of the process. For NCS Pearson, this was simply not an issue. They were extremely confident that they have the quality assurance techniques and recovery methods to eliminate any potential problems.

NCS Pearson did admit to some problems with the scanning, indicating that a greater proportion had to be individually scanned than they had expected.

4.2.4 Recruiting and training markers

The markers involved in the Pilot were not of the calibre intended. Few had been markers before, many had limited IT experience and some were teaching in (educational) areas far removed from Year 7. AQA were to recruit markers, but NCS Pearson felt they were not helpful. AQA blamed the situation on the lateness of the process (delayed until the Pilot was agreed) and on initial instructions that a national sample of markers who would be accommodated in Hellaby, being countermanded by a request to find teachers local to Rotherham.

Again the lesson is that work and responsibility must be agreed early, even in a pilot. QCA was well aware of the continuing issue of agencies, forced to work together by QCA, not doing this willingly or well, but felt it had procedures to deal with this.

4.2.5 Software functioning

The software was not much commented upon and, according to the participants, appeared to have done as it was supposed to. However, one "glitch" was identified. Only NCS Pearson alluded to this, but it has since been reported to the Project Board. This was that the seeding of previously marked items by the Chief Marker back to individual markers was not possible. According to NCS Pearson this was not part of the software obtained from China, so they "retro-fitted" this capability and it did not operate successfully. They seemed not to regard this as a major matter and asserted that it would work in the full ePEN system.

4.2.6 Operating the marking process

NCS Pearson regarded this as a success – the markers completed all the scripts available and worked faster than expected. The limited number of scripts prevented a full trial of the speed of working and therefore capacity. English markers continued to use the paper-based mark scheme rather than the on-line one. This may have derived from the conventional training instructions.

NCS Pearson did, though, consider that the reporting aspects of their software needed improvement. As they stood, they gave very detailed information on small parts of the process rather than useful overall measures.

4.2.7 Reporting issues

The proposed reporting method of putting script images on a CD to return to schools (actually to QCA in place of the schools) was found not to be practicable. For larger schools several CDs would be required, which would make it difficult to handle statistical data reporting. There was some incredulity from AQA that the necessary calculations on capacity had not been done. QCA too felt that this should have been predicted. There are now suggestions that data will be returned over the internet and NCS Pearson are exploring this. However, capacity problems also apply to this method.

Several of the contributors regretted that the return of data to schools was not being given a higher profile in the Pilot.

4.3 The business case

The business case set out by QCA (and its equivalents in Wales, Scotland and Northern Ireland) "focuses on providing a more flexible service, a more automated service so reducing paper burdens, a more seamless approach across all parties involved in the testing and marking process and a better value service." Within this, five groups of stakeholders were identified for the proposed improvements. These were: markers, teachers, local education authorities, government departments and their agencies and Ofsted. The benefits in relation to each were directly addressed in the interviews and are set out separately in the following sections.

In the second set of interviews, the general success of the Pilot (albeit with limited information yet available) had led to several correspondents believing that the business case had already been strengthened.

4.3.1 Marking by non-teachers

In the current pilot, questions identified as 'simpler' were not marked by teachers; they involved either responses entered manually and scored by computer or responses marked by clerical markers.

There was general agreement that this is desirable and valuable for two reasons. First, it releases teachers for the more demanding questions, which require their professional capabilities. Second, but related to the first, it may assist with recruitment of markers – currently an area of some difficulty.

Two cautions were provided. First, it may be that the use of clerical markers is not publicly acceptable and could not be countenanced, for example at A-Level. Their use may be perceived as a lowering of standards. (NCS Pearson suggested that greater cost savings could be made by taking this element of marking overseas – but even they did not think this would be politically acceptable.)

The second caution was more of a statement of opportunity – with suitable changes to the tests, clerical markers would not be needed since all these objective questions could be marked electronically, either by Optical Mark Reading or Optical Character Recognition (though some doubts were expressed over current software's ability to read children's writing).

4.3.2 Increased throughput of marking, leading to better remuneration to teachers

Although all the respondents expected to see an increase in the speed of the marking process, they did not regard it as important that this led to greater payment for teacher markers. It was acknowledged that different remuneration schemes will be required for on-screen marking, perhaps piecework with bonuses for quality, but this element is yet to be worked out. It is certain that payment per script will not be appropriate and that a completely different model of payment is required.

4.3.3 Removal of clerical tasks

This is universally seen as a major benefit. Comments ranged from wondering why it has not happened already to the high extent of inaccuracies of the current markers. Even if only clerical reviews were removed, this would solve a large problem for QCA and the EMAs. The only note of caution was a suggestion that there should be double checks in the software to ensure that there are no inaccuracies in the totalling routines.

4.3.4 More detailed performance information for schools and teachers

Again, this was generally agreed to be a valuable aim. Surprisingly, the most in favour were NCS Pearson, although they probably had a notion of great amounts of data being returned to schools for them to examine themselves. Others were more aware of the overload on schools and expressed caution in the way data are to be provided, which needs to be easily interpretable. For DfES and QCA any such provision of data must contribute to teaching and learning in schools.

DfES was insistent that any such data use must fit into its Information Management Strategy, allowing integration into the Common Basic Data Set. This integration must not cause work for teachers or schools.

A long-term possibility suggested by two participants was the use of an extranet, allowing schools password-guarded access to check on the progress of their own results during the marking process.

4.3.5 Increased accuracy of marking

All were agreed that this was a valuable aim and one that should be met from on-screen marking. The facilities for double-marking, for checking on markers right through the process, plus the specialisation in fewer questions should all lead to greater consistency. This should then reduce the rate of requests for reviews. Two comments reflected some caution. The DfES needs to ensure that stakeholders in the system have confidence in it and, again, the issue of non-teachers marking and its effect on confidence was raised.

In part, the increased accuracy derived from the double marking or the expectation of it. In this pilot, only around 10% of 'expert items' were marked by two different markers. Increased accuracy could be obtained with 100% double marking but this would have significantly higher costs.

4.3.6 More data for government agencies

This was not regarded as important by all. There was a recognition that the collated data could provide some curriculum analysis which might be used in a general sense for monitoring or as feedback to LEAs and others to inform training. A further suggestion was that it could assist the test development process (although it seems a bit late for that function). Some of the interesting analyses would require the data to be linked with background pupil data from other databases, which might be technically and legally difficult. As mentioned above the DfES is eager to ensure that the formats used conform

to government recommendations (Common Basic Data Sets) so that linking to other data sets in education is facilitated.

4.3.7 Reductions in time – increasing the speed of the process, returning results to schools earlier

This was generally regarded as important and a possibility. NCS Pearson were very bullish – expecting to be able to reduce the whole process to four weeks (currently the process takes four weeks just for the marking). Others did not see such large reductions as a possibility but did expect some speeding up. The DfES was especially interested in the ability of future systems to return data to schools by the end of term.

4.3.8 Reductions in costs

There was a sharp divide between the government agencies and the current operators on the one hand and NCS Pearson on the other. NCS Pearson saw that great savings could be made, of the order of 20%, depending on the remuneration system employed. NCS Pearson also made the point that really large savings could only come from changes to the test format or other radical proposals such as moving the marking abroad where labour is cheaper. Other priorities would make this impossible.

Generally most saw the more likely scenario being that costs remain the same, but that value is added through extra speed, greater reliability, more information for more uses etc.

The present pilot was not seen as providing sufficient information to model costs upwards, since the scale was too limited.

4.4 Scaling up

4.4.1 How well will the process scale-up?

There was a general expectation that on-line marking will come into use for national curriculum tests at some time in the future, though not immediately. All expected that there would be a further stage of development for the Year 7 progress tests before moving to national curriculum tests. In the later interviews, there was a growing acceptance of the need to introduce larger numbers and a wider range of ability and subjects. The interviews gave very little information on how well the process would scale-up, beyond thoughts and commentaries on the issues to be faced in the process.

4.4.2 Possible constraints and risks in scaling up

The possible constraints and risks to scaling up which were identified could be classified into four groups: the technology, people, the process and assessment issues.

4.4.2.1 The risks from technology

Put bluntly, as one respondent did, the risk is that it will not work. This did not seem to be so from the Pilot, where the basic process operated. However, it was not totally perfect. One element, the sending of 'seed items', did not operate because of a software "glitch". In part this was because a simpler level of software from China was used by NCS Pearson, rather than the full US system. The decision to do this was taken because of limited lead in time. Risks increase from speed, emphasising the need for early decisions and time for testing. Even when testing of systems is undertaken, late changes or tiny unforeseen situations can have untoward effects.

In addition to the software, there are risks from the hardware, particularly if a devolved solution is adopted with either home-working or the use of centres in schools. Would all the systems be capable of running the software at sufficient speeds? Would sufficient support mechanisms be able to be provided? Since the internet would be utilised, would it be capable of carrying the traffic speedily? Would there be additional risks from viruses (general or specifically targeted) or from hackers?

Some of these risks would not arise in any pilot and could not be realised and evaluated until there was an operative system.

The largest risk is likely to arise from increases in volume. The present pilot is very small scale and even if extended to all Year 7 progress tests would still not represent the size of even a single national curriculum test. It seems generally agreed that there must be a further pilot with much greater volume and some pressure on the system. In many government IT developments, systems operate well at the trial stage but then fail or reveal problems when they come into mass use.

4.4.2.2 Risks from people

A main identified risk to the system was the provision and compliance of the markers. Markers, in any case, are in short supply and the most difficult marking area, English, is the hardest to recruit for. The external marking agency was not particularly successful in recruiting experienced markers for NCS Pearson for the Pilot and this required only a very small number. Hence, the pool of existing markers (and perhaps new markers) must be

convinced of the attractiveness of marking on computer, both in terms of the manner of working and remuneration.

Whereas markers can be transferred from the existing system there will need to be a new set of experts and managers to run the computerised marking system. These will include project managers, systems experts and IT support specialists. At present neither QCA nor NCS Pearson has sufficient staff to fulfil its expected (or desired) role in the full operation. The current operators have staff but no disclosed expertise in the area. Generally, there was a wariness of involving new companies which do not have knowledge of the educational scene. For these reasons, the programme for moving to full implementation will need to take account of stages for the development and training of the staff required at each stage.

4.4.2.3 Risks from the process

The management of the process of scaling up may pose a risk. Although individual projects or development may be managed well using PRINCE2 methodology, other factors may intrude. One identified by the Pilot was the intrusion and interaction with other innovations. In the case of the Pilot, confusion was caused by the introduction of the Year 7 optional tests at the same time and the instructions to some schools (in the KS3 strategy pilot) to use those tests. It is unlikely that there will be no other innovations and pilots either in national curriculum testing or more broadly at the time of scaling up. These interactions will need considering and these effects estimated and ameliorated.

A particular instance of interaction may be the DfES's desire for data to fit within its own Common Basic Data Set. If this is to be the case, it should be specified from the outset and the manner of the integration should be clear.

Outside of QCA/DfES there was a general view that decisions to proceed are taken late and implementation then hurried. There is almost an expectation that contractors will have to make up the time that is lost by QCA/DfES. This is regarded as unfair, but more importantly, the short lead-in times and hurried implementation do pose risks. QCA itself is aware of this, but has not been able to break out of this cycle.

The current QCA/DfES preference is to divide systems into "chunks" and use several contractors. The contractors themselves see this as inefficient, requiring separate databases and the repeated passing of information sometimes between semi-compatible systems. There are risks from too many contractors because each interface is a risk either

technically or in personal/organisational terms. Alternatively there are risks in placing too much of the whole system with a single company, if that were to fail.

There was a common view that introducing agencies without a track record in education, simply working from a QCA specification and offering the lowest price, would again lead to heightened risks of failure.

4.4.2.4 Risks to assessment

The first set of risks to assessment derive from confidence in the process. There was a strong view that this derives from the expertise of the markers (although other public sources emphasise the poor quality of the marking and the large number of requests for review). There could be risks to credibility if the role of markers were diminished or removed completely, for example, if computer algorithms alone were used to mark essays. The level of marking at which such confidence is lost remains a matter of conjecture.

There are risks to the validity of the assessment, if the marking process begins to drive the nature of the assessment. Generally speaking, the style and content of the national curriculum tests commands broad support. This could be reduced if the form of testing were amended, particularly in directions which produced large cost savings.

Finally, the change of marking process may indirectly lead to changes in the resultant test standards. For example, because of the separate marking of individual questions, the marking could become demonstrably more reliable. However, a subsequent removal of borderlining (due to this increase in marking reliability) could end the effective boosting of marks around the cut score. This could cause a discontinuity with previous test standards. In essence, greater marker reliability may not necessarily imply a consistency of standards with the past.

4.4.3 Issues to be addressed in scaling up

The interviews threw up a substantial number of issues which will have to be addressed in the process of scaling up. Many of these would be rapidly disposed of, if a particular solution were to be operated. Others become more pressing with particular solutions. The issues have been collected into seven categories, though many interact and cross into other groupings.

4.4.3.1 Implementation timetable

There are a number of models for the scaling up timetable. The interviewees proposed differing means of proceeding. There was initially a universal view that there should be a further year of development with the Year 7 progress tests. This should move towards a full scale implementation to ensure that systems could cope with high volumes. It would also allow a further opportunity to pilot the return of results to schools, gauging the usefulness of this feedback.

In the second interviews, some views had shifted and there was more support for exploring new areas, including partial marking of some key stage tests. The main reasons for this appeared to be in order to move forward with tests which had a full ability range and which allowed a greater test of throughput. There was also a desire to explore home-based marking.

Beyond this next stage, viewpoints differed markedly as to the medium term. Some advocated a start on a single national curriculum test (say KS2 mathematics) others a whole key stage, others starting with several tests but only part of the volume for each. It will be important to come to a view on the expected timetable and to publish and promulgate it. Those involved felt that this would allow a clearer view of the process and a lead-in time which allowed planning. In contrast, annual decisions for the following year did not allow planned development; with longer contractual periods the viability of extensive investment and training is improved.

Some specific issues need resolution as part of the timetable. First, the length of time that parallel processing (conventional versus new technologies) will be required. The view of NCS Pearson is that if the current pilot shows the equivalence of marking for conventional and on-screen methods, there would be no point to further parallel operation. In contrast the view from the EMA was that parallel working should be maintained for as long as possible to give absolute security (see the discussion on risk below). A further point made about parallel processing was that it is undesirable because it leads to dilemmas as to which result to accept where they differ. Solving these adds to the complexity of the process and the time taken.

While most of the interviewees saw that there should be a gradual increase in the amount of on-screen marking, building up to all NC tests, one considered that the exclusion of certain subjects in the initial stages would send untoward messages about the status of those subjects.

4.4.3.2 The marking system

Most respondents still had an open mind on the eventual system and how it would operate. As a result of the Pilot so far, NCS Pearson have revised their views from believing that there would be a home-based system with markers operating on their own, to believing that there will be a centre-based operation. Part of the reason for this was that the markers themselves preferred human contact and the immediacy of access to the Chief Markers. These reasons are not, of themselves, recommendations for on-screen marking, since they apply equally to centralising conventional marking, which is being tried by AQA this year, apparently with some success.

The provision of marking centres poses difficulties since marking national curriculum tests is concentrated in time at one point during the year. It would therefore be extremely expensive and wasteful to have centres and computer equipment dedicated to this alone. The most commonly suggested alternative was that the IT centres of LEAs or schools should be used, over half term and during weekend and evenings. This would lead to a relatively devolved operation reliant on many separate schools and LEAs with considerable contractual negotiation and organisational requirements. It is likely that some of these would fail each year, but given the nature of the marking process, these elements could be diverted elsewhere.

It was noted that the alternative of a few large centres might attract large costs associated with travel and (particularly) accommodation for markers. Such expenses can be greater than salary costs. The regional centres will need to be situated to minimise travel and subsistence costs.

The only alternative suggestion to educational centres was to utilise the network of driving test centres which are computerised and run by a single assessment organisation. This would have greater coherence and simplicity but would limit the hours of marking to evenings and Sundays.

The alternative model was for marking to be undertaken at home on the markers' own computer equipment. (One respondent assumed that equipment would need to be supplied to all markers for this purpose and began to calculate a very large cost figure.) Home working throws up a new set of issues. The first is of course the software and whether it will operate on many machines of quite diverse specifications. NCS Pearson did not regard this as a problem since their software uses browsers and they believed that any specific software required could simply be automatically downloaded. Others were much more sceptical that older machines would cope.

Various aspects of legislation were referred to. Health and Safety legislation requires proper workstations for workers and display screen equipment checks, perhaps including eyesight examinations. Would these be done or enforced? It also requires reasonable hours to be worked at the keyboard/screen with adequate breaks. Could this be enforced? It is known that currently many markers work long hours and late at night. Would this be permissible? A second area of concern, for some, was the Data Protection Act. Since browsers automatically store pages of information, home machines might retain confidential information on individual pupils. The systems used would need to ensure that this did not happen.

Finally, there was the issue of contractual responsibility for damage to home machines or other software caused by the marking software. Would liability be accepted?

IT and software support for home-working would also be necessary, with help lines or on-line help built into the system. The structures and management of these would need establishing. There seemed to be an assumption that on-line home marking would be cheaper than current practice. This was challenged by one contributor, who pointed out that it depended crucially on the existing computer equipment of markers and whether they would have to be bought high specification machines. If this were the case, then savings might be very long term.

One cautious contributor expressed doubts about the principle of on-screen marking at home because (with current technology) the contractor could not know who was actually doing the marking. With conventional marking there is a paper-trail of written information and verifying signatures but this would be lost. (It is the case of course, that the enhanced reliability checks of the on-screen marking ought to detect substitute markers, provided they were not consistent or accurate.)

A very general issue affecting both home-marking, and probably marking in centres in schools, is the cost of internet access and the level of broad-band access in Britain. These systemic factors may limit the possibilities for devolved marking.

Overall, there was a strong feeling that this option needs investigation in whatever is decided for 2002. In general, contributors were arriving at the conclusion that the eventual system will have both centre-based and home-based marking to ease the supply of markers and yet secure the advantages of screen-based control.

4.4.3.3 Public acceptability

There are risks to the system in terms of parental and public acceptability. The perceived use of machines to mark children's work could be regarded as dehumanising and rigid, not giving the children a fair chance or human understanding. In addition, failures even for the results for few children would be blown up into IT horror stories by the press. Finally, risks to confidentiality could arise. These dangers would need to be handled through good communication, good piloting, learning the lessons required and building confidence through good management. There is likely to be scepticism (public and teacher) and this will need to be overcome through publicising the advantages of the system, especially increased reliability.

4.4.3.4 Scanning

Prior to the marking process the scripts would need to be scanned. This could, in theory, be undertaken at a single centre operating with a large capacity. One of NCS Pearson's respondents advocate this. The other respondent, and most interviewees, thought there should be a regional system, partly to minimise the risk of disasters but also to spread the distribution.

In the light of concerns from some correspondents about the potential for data loss in the scanning process, this aspect will need to have careful monitoring in future pilots and the establishment of public quality assurance procedures and acceptability criteria.

4.4.3.5 The markers

The point was made that the markers are currently in a strong employment position. There is a shortage of qualified personnel who are prepared to undertake the work. The movement of the marking process to on-screen marking cannot happen without the group of markers being willing and able to make it. Hence the attitudes of markers to the process and also marker behaviour must be monitored at each stage of development.

Beyond this, the use of teacher markers needs careful consideration. The process in the Pilot, of separating some questions to be marked by clerical markers, will reduce the reliance on teachers. This reliance could be further reduced through changes to the tests (see 4.4.3.7). The issue of public (and professional) acceptability was raised by the EMA and NDCA respondents. Their view was that confidence in the marking system derived from the professional input of teachers and there were dangers in reducing this. Generally, participants wished to retain teacher involvement where it was professionally required, but saw advantages in clerical workers taking over aspects which did not require

teacher involvement. NCS Pearson noted that further savings could be made if the data were transferred abroad, allowing cheaper labour costs. The public acceptability of such a development is likely to be low, and, if such a view is taken, it needs to be overtly ruled out.

It was evident that new pay structures need to be established for on-screen marking. The previous practice of piece-rates based on scripts is not appropriate. Payment by time will not reward high-quality rapid marking. It would be desirable for markers to be paid according to the number of questions marked but some questions are likely to be easier and quicker to mark than others.

All this suggests that the data from the present pilot should be used to develop and examine a new reward mechanism to be trialled in a subsequent pilot. The mechanism might incorporate both speed and marker reliability as measured by the quality assurance procedures.

An extension to the present proposals for further work would be to develop means of training markers on-screen, as well as perhaps the processes of recruiting, qualification and training. These might attract an extra pool of markers, easing recruitment difficulties.

4.4.3.6 Managing the system

All the respondents assumed that QCA would not operate the marking system themselves but would contract it out. It also seemed to be assumed that, as at present, the whole process would be divided between several contractors, each acting sequentially, or in some cases, in parallel (e.g. there might be several marking agencies). This strategy appears to be for two purposes. First, to minimise the risk of complete failure, and second to create or maintain a market in which there are a range of suppliers. Those in receipt of QCA contracts pointed to the inefficiencies of such systems: the delays introduced by the interfaces between agencies; the inefficiency of several databases rather than one; and the costs of passing data between these databases. This was contrasted with the work of examination boards where all functions are under one management. A regional system is a possibility, but this can lead to different systems, styles, procedures and documentation. A split between subjects may be a more standardised alternative.

There is a danger that the rush to utilise NCS Pearson's expertise in technology may undervalue the existing expertise of AQA (and others) in managing markers and the marking process in a fair and reliable manner. There are many elements of the "traditional" system, such as recruitment, training, standardisation, contracts with markers

and relationships with schools, which must be incorporated into the new system and it would be foolish to discard the existing expertise without careful consideration.

These considerations may have to be related to the capacity of QCA itself to manage a changing system using its current staff and resources. The DfES now seems content with QCA's project management skills, but there is no doubt that the effort of managing several agencies (and causing them to work together in the desired way) may add to the resourcing requirements. QCA will need to devise a strategy for the management of the system which is within its own resources and which involves the optimal number of agencies for reducing risk yet providing efficiency gains. In addition, QCA needs to be better at planning, setting up contracts earlier and smoothing implementation. This is acknowledged but seldom achieved.

A project management style that allows a strong relationship with contractors so that problems and changes are notified and discussed is also seen as essential.

4.4.3.7 The tests

There is no doubt that the marking process could be automated to a greater extent if the nature of the tests were changed. At one extreme, completely multiple choice tests could be quickly and reliably entered and scored without expensive professional input. However, this was generally considered to be educationally undesirable and QCA's Board would not countenance it. Smaller changes to the tests would allow some questions to be marked automatically through the scanning of particular areas for marks and lines. Yet, such systems would not have a human marker's capacity to detect the intentions of a child who indicates a correct answer but uses the wrong answer mode. At the next level, a greater proportion of questions could be written and included for clerical, rather than professional, markers. Ultimately it may be that children's writing of numbers, letters or single words can be read by software with sufficient accuracy for clerical markers not to be needed. As this illustrates, there are many shades of possibility and many views on whether each is educationally acceptable and valid in the context of the assessment.

In the present pilot, no changes were made to the questions since these were to be the same as for other pupils and the effect of changes could not be known. The lead-in time for test development is two years from specification to the use of the tests. The timetable for scaling up therefore needs to allow for changes to the tests to be made at the specification stage so that they can be incorporated during development. This implies that the parameters of any such changes need to be established well in advance, following debate between all the interested parties.

In part this is also related to the purposes of introducing electronic marking. If this is to save large sums of money, then that can be done, but the tests would need to change. The extent to which economics are required and the limitations put on changes to the tests need stating.

4.4.3.8 Confidence and risk

There were a number of attitudes to risk shown in the initial interviews. The EMA and NDCA representatives (from within the traditional examination board setting) seemed to expect that a full implementation with no errors would be required. They were also concerned that both QCA and DfES had conflicting requirements, wanting innovation but no risk. Indeed, the DfES commented that they would advise ministers not to proceed if there were any risks. In contrast, NCS Pearson were prepared for problems during implementation and saw this as a natural part of a change process. This is not to say that they would not implement proper project management and control.

The views of the EMA/NDCA (each from within an examination board) appeared to arise from a laudable public examining mentality of accuracy for all individuals' results and maintenance of public and professional confidence. These aspects are crucial to (say) A-level but a view could be held that they may not be so important for national curriculum tests, where errors could be subsequently corrected with less disruption. Alternatively the viewpoint might be that national curriculum test results must be assured of being absolutely accurate. The view taken of risk will contribute to determining the timetable and processes for scaling up the current pilot. A cautious view would imply a long timetable with parallel working with conventional methods and elaborate safety and rescue procedures. This would increase costs, reducing one motivation for implementation. A less cautious view might allow more rapid implementation and an earlier prospect of benefits.

These issues were explored more explicitly in the second interviews and a greater concordance of view emerged. This recognised that there could be no risk-free options and that as a consequence, risk needed managing properly through contingency planning and sound project management. The development needs to be planned as an interactive growth process which may well have on-line and conventional systems alongside each other for a period of time. Overall a cautious but not stifling approach was advocated.

4.5 Conclusions

The six participants brought to the interviews the expectations and attitudes of their own organisations and their own role in the Pilot. This led to some differences in viewpoint and emphasis. However, these were not great and the expectations for the Pilot were fairly concordant. Generally speaking, all acknowledged that it was a limited exercise but thought it would be successful in its own terms. The greatest limitation, expressed by all was the small volume of marking processed. This was less even than had been planned, due both to conflicts at an organisational level and the needs of the parallel marking process. It is imperative that there is a further pilot with much more stress on the actual demands of marking volume, and total completion for all scripts.

Having said that, there was considerable agreement that the process of development should continue. The current pilot was regarded as a springboard for further work. All said that the Year 7 progress tests provided a good route for this since they are important but not statutory. There should be a further pilot in 2002, possibly including all the Year 7 progress tests, but certainly a much greater volume of scripts. Further extensions which could be included are marking at home and the return of results to schools, including diagnostic information.

The process of scaling up to national curriculum tests will need careful design. There remain several possible models, including centre-based and home-based working (or both). Many other issues follow from this and must remain open until a more definite model is decided upon. It would be helpful though to have a provisional timetable as soon as possible, setting out the expected years in which each test is likely to move to electronic marking. This would allow planning for staff procurement and development as well as logistical and infrastructure planning.

This brief survey has illuminated some fundamental issues for the scaling up process, which are to do with its underlying philosophy rather than the technology. These are concerned with the attitude to risk, the structures of contracting, the extent of changes to the tests and the management of the process. These should be discussed in agreeing the programme of implementation and its parameters, so that they are known to all concerned.

Section 5 A statistical analysis of the Pilot

One of the potential benefits of implementing new technology for the management of national curriculum test marking is the provision of substantially more statistical and management information on the process. Studies 4 and 5 collected quantitative data from the Pilot alongside similar data from the conventional marking processes.

The studies began with the construction of a *Specification for Data Collection and Analysis*, prepared by the NFER in collaboration with NCS Pearson and QCA. This outlined the kind of management and measurement data that would be required to complete the exercise. The intention was to collate the most relevant and potentially insightful data that arose from the Pilot and to compare them against similar data from a range of conventional marking systems in the UK (predominantly recent Year 7 progress tests, key stage 2 tests and, to a lesser extent, key stage 3 tests).

In fact, it was not always possible for QCA to locate similar data relating to conventional marking processes. Therefore, the comparisons of data between new technology marking and conventional marking were not always direct. As already mentioned, one of the potential benefits of new marking technology is the straightforward provision of data that is not routinely being collected at present.

The following analyses served three main goals:

1. to illustrate the variety of management and measurement data that it was possible to collect using the technology of the Pilot;[1]

2. to provide baseline data on new technology marking for comparison with future pilots and trials;

3. to compare data arising from the Pilot with data arising from conventional marking and, therefore, to gain some indication of how successful the Pilot was.

It should be recognised that many caveats need to be taken into account when interpreting the data presented below. This is particularly true with respect to comparisons made between systems. Where possible, important caveats have been explicitly stated. The

[1] Note that this is only an illustration and the Evaluation does not present information on the full range of management or measurement data that arose.

results of studies 4 and 5 are presented in two separate sections, corresponding to the management and measurement data respectively.

5.1 Analysis of the management data

Study 4 focused primarily upon management data, where this was taken to be quantitative information on the speed and accuracy of procedures for the processing of Y7PT scripts through the marking system. The central question of study 4 was whether the implementation of new marking technology could be expected to eliminate inevitable procedural errors and, more generally, to speed up the marking process. These two issues, of accuracy and speed, are not independent (as, for example, more accurate processing is likely to result in less delay).

The data underlying the following analyses are presented in Appendix 5.1. They were provided for the NFER by NCS Pearson on 6 August. (The tables in Appendix 5.1 are not exactly as provided, as a number of presentational and statistical modifications have been made).

5.1.1 Processing speed and load

The first issue considered was the extent to which procedures of the Pilot either speeded up or slowed down the processing of scripts through the system. Evidence concerning the time taken to complete various stages of the Pilot procedures was collated.

5.1.1.1 Processing script batches through the scanners

Question: ***once scripts had been sent to NCS Pearson, and logged in as batches, how much time elapsed before each batch was guillotined and scanned into the system?***

The intention of these analyses was to establish the amount of time required by the initial stages, prior to the commencement of marking. One of the particular concerns of the Evaluation was to determine whether the central technology of the Pilot – the scanning process – would, or could, constitute a potential 'bottleneck'.

Table 5.1.1A (Appendix 5.1) presents information on the time lag between the end of processing of a batch of scripts on DWS Batch Builder and the commencement of

processing of the same batch on DWS Scan Master.[2] This includes time taken to guillotine each script within a batch (removing the spines prior to scanning).

The fact that the standard deviations of the time lags were larger than the means demonstrates that there was considerable variation in time lag across batches (for each of the six papers). This was supported by an analysis of the minimum and maximum time lags: across papers the smallest lags ranged between 2 minutes and 16 minutes, while the largest ranged between 3.9 days and 6.0 days. The largest time lags should be interpreted with the understanding that some batches were delayed across weekends when no scanning occurred. On average, the time lag between post-receipt batch processing and scanning ranged from 12.8 hours (English reading) to 22.1 hours (maths A).

The principal cause of delay at the scanning stage occurred when scripts could not be scanned automatically, due to a range of problems (see 2.1.6 and 2.2.6). Those scripts then had to be sent to a flat-bed scanner and processed through DWS Scan Master Exceptions. Here, each 'rogue' script was scanned in its entirety. Significantly, even if there was only one 'rogue' in a batch, the entire batch was held up. Thus, a large number of exceptions would not be desirable.

Table 5.1.2A indicates that the number of 'rogue' scripts ranged from 40 to 143 across papers. The percentage of 'rogue' scripts clustered between 1.2% (mental arithmetic) and 1.7% (English spelling and handwriting), with English writing being something of an exceptional case at 5.1%.

It is clear from Table 5.1.3A that many batches were held up as scripts waited to pass through Scan Master Exceptions. In fact, 72% of batches were held up for English writing while even for mental arithmetic, the least delayed paper, 22% of batches were held up. Comparing the number of papers (5.1.2A) and the number of batches (5.1.3A), it seems that there were typically only one or two 'rogue' scripts per batch. Particularly considering the high mean time lag between processing a batch on Scan Master and processing it on Scan Master Exceptions (which ranged between 42.6 hours and 71.5 hours), the frequency of scanning errors might constitute a threat to the timely processing of test scripts. On the other hand, NCS Pearson claimed that the level of exceptions was higher than anticipated and higher than would be expected in a full scale-up. They proposed that modifications to test paper design could help prevent many of the problems

[2] The suffix A (as in Table 5.1.1A) indicates that it is located in the Appendices.

that occurred (although design modification would not necessarily prevent pupils from writing in places that they were not supposed to).

Table 5.1.4A presents information on scanning rates, for each of the six papers, for the main high speed scanner. It records average scanning rates ranging between 492 sheets per hour (English writing) and 2,058 sheets per hour (English reading). Quickest scanning rates ranged between 2,558 sheets per hour (English writing) and 4,000 sheets per hour (maths A).

None of these rates was up to the original planned rate of 5,760 sheets per hour. However, it was explained by NCS Pearson that the slower rates were a consequence of the requirement to scan each script for both a full-page image and for item clip images. Overall, the throughput of the Pilot was 2,880 sheets per hour. The corresponding throughput for the Exceptions process was 15 sheets per hour and, for the Attachment process, was 30 sheets per hour.

Finally, Table 5.1.5A presents some of the reasons why scripts were rejected by the high speed scanner (see also 2.1.6.2). These typically involved pupils writing in areas that they should not have done (i.e., obscuring pre-printed information necessary for effective scanning). Unfortunately, no information was available on the relative frequency of these causes.

5.1.1.2 Marking of scripts

Question: how long did it take for markers to mark item responses?

Turning to the marking of scripts, the software provided ample opportunity for the provision of management data because it automatically recorded the time that a marker was logged on to an item for marking. Indeed, these 'productivity data' were available to administrators daily through the use of specific reports.[3,4]

Tables 5.2.1A, 5.2.2A and 5.2.3A present productivity data for response selection markers, clerical markers and expert markers, respectively. Data from these tables are

---

[3] These reports are not considered further, as the present analyses focus exclusively upon full aggregate data.

[4] The marking time data should be treated with a certain amount of caution as there may have been occasions when markers were presented with an item but they were not in a position to mark it immediately.

summarised below, in Table 5.1.[5] Care should be taken in interpreting the mean (and median) figures as large standard deviations betrayed considerable (within-question) variations in the time taken by markers to mark item clips.

Table 5.1 The median (across questions) of the mean number of seconds taken to mark individual clips.

	Response Selection		Clerical		Expert	
	Median Secs	(No. Qns)	Median Secs	(No. Qns)	Median Secs	(No. Qns)
Reading	6	(1)	5	(8)	15	(21)
Writing	-	-	-	-	367	(1)
Spelling/Handwriting	-	-	2	(20)	80	(1)
Maths A	3	(16)	3	(11)	9	(7)
Maths B	3	(9)	3	(15)	6	(10)
Maths Arithmetic	1	(15)	1	(5)	-	-

Table 5.1 provides a rough indication of the amount of time taken to mark different types of question across the six papers. The indication is rough, to the extent that they were averaged across markers (mean – within questions) before being averaged across questions (median – within papers), thereby ignoring the substantial variance in the data. The data were also rough to the extent that different markers marked different numbers of clips for each question. However, it seems likely that the data still provide a reasonably useful indication of the amount of time that it typically took to mark item clips.

The trends in Table 5.1 are reasonably clear; response selection markers and clerical markers took a similar amount of time to mark item clips, while expert markers tended to take around two to three times longer. This is understandable, as the expert marker questions presumably required more attention. Arithmetic questions took the least time to mark, requiring only a second on average. Writing responses took longest to mark, taking over six minutes on average.

---

[5] From Tables 5.2.1A to 5.2.3A, it will be noticed that the first question for each paper tends to have a longer mean time per item than the other questions do. This query was raised with NCS Pearson who responded as follows: "This is a factor of software design. The audit timestamp is recorded when a document is locked to an operator for test paper within the editing process. There may be a delay of fetching, decompressing and displaying the image before the operator can start to do anything. In hindsight we should move this audit action to prevent this from occurring again."

By multiplying the number of questions in each group by the median time taken to mark them, we can arrive at a very rough estimate of the notional average time that it took to mark each script on-line.[6] These values are presented below in Table 5.2.

Table 5.2 A very rough estimate of the notional average time taken to mark each script on-line during the Pilot.

	seconds	minutes
Reading	357.0	6.0
Writing	366.5	6.1
Spelling/Handwriting	120.0	2.0
Maths A	144.0	2.4
Maths B	132.0	2.2
Maths Arithmetic	20.0	0.3

Tables 5.2.1A to 5.2.3A present marking speed data from the Pilot, aggregated across markers but separately by question. This kind of question-level data is likely to prove very useful for senior markers and researchers. Investigations into the characteristics of questions that take longer to mark than others is likely to yield information that may be fed back into the marker training process and, potentially, into the test development process.

These data on marking speed can also be produced at a dis-aggregated level and presented separately by marker, on a daily basis, as the marking is taking place. Such data can support the identification of markers who are taking undue amounts of time to mark items. However, it will be crucial for future pilots and trials to determine how this kind of information can be used most effectively. Not only are they the kind of data that invite over-simplistic conclusions; even if interpreted validly, they might lead to negative consequences. For example, if the data were used simply to draw attention to slow markers then there would be a risk that marking would tend to speed up at the expense of rigour.

[6] Multiplication of the median by the number of questions has been used simply because of the point mentioned in the preceding footnote. However, if the mean marks are added across questions, a similar result obtains.

5.1.1.3 Supervision

Question: ***what supervisory demands were made upon the supervising markers?***

Clerical and response selection markers could refrain from marking a particular response that they were unsure about by sending it with a sticky note directly to a supervisor. Table 5.3 is a summary of data presented in 5.3.1A and 5.3.2A (which capture the same information but for individual questions of each paper). It is clear from this table that the sticky note function was hardly used for English. It was used more frequently for maths, where the highest median figure (74 clips) reflected 1% of the total number of clips viewed.

Table 5.3 **The median (across questions) of the total number of clips sent to supervisors by clerical and response selection markers.**

	Response Selection			Clerical		
	Median No. Clips	Median % Clips	(No. Qns)	Median No. Clips	Median % Clips	(No. Qns)
Reading	7	0.2	(1)	4	0.1	(8)
Writing	-	-	-	-	-	-
Spelling	-	-	-	5	0.1	(20)
Maths A	69	0.9	(16)	36	0.5	(11)
Maths B	74	1.0	(9)	24	0.3	(15)
Maths Arithmetic	67	0.9	(15)	5	0.1	(5)

Unfortunately, these figures do not tell the whole tale of supervision, as the on-line system was not entirely effective and markers typically requested advice from senior markers when they were physically present.

Expert markers were not able to refrain from providing a mark for each clip. However, it was possible for them to send a supervisor a message regarding a particular item (explaining any concerns that they may have had with the mark that they had awarded). Table 5.3.3A records the number of messages that were sent. Clearly, the number of messages sent was very small, particularly for English, despite there being many questions for expert markers to mark. The implication is that this function largely failed as markers preferred to gain advice directly and in person.

5.1.1.4 Adjudication

Question: ***what adjudication demands were made upon the supervising markers?***

For all clerical and response selection questions and for a sample of expert questions, two marks were provided for each item. Hence, it was necessary for a senior marker to stand

in judgement when any of the mark-pairs were in disagreement. This process was known as adjudication. The data for the adjudication of response selection, clerical and expert markers is presented in Tables 5.4.1A, 5.4.2A and 5.4.3A, respectively. The data are summarised in Table 5.4, below, which shows the median of the total number of items that required adjudication across questions.

Table 5.4 The median (across questions) of the total number of items that required adjudication.

	Response Selection			Clerical			Expert		
	Med. No.	Med. %.	(No. Qns)	Med. No.	Med. %.	(No. Qns)	Med. No.	Med. %.	(No. Qns)
Reading	95	4.5	(1)	25	1.2	(8)	44	19.3	(21)
Writing	-	-	-	-	-	-	n/a	-	(4)
Spelling/Handwriting	-	-	-	50	2.5	(20)	131	58.5	(1)
Maths A	542	14.3	(16)	119	3.1	(11)	47	13.0	(7)
Maths B	387	10.3	(9)	73	2.0	(15)	37	10.1	(10)
Maths Arithmetic	476	12.7	(15)	52	1.4	(5)	-	-	-

For clerical and response selection markers, the median of the total number of items adjudicated ranged from 25 (English reading, clerical) to 542 (maths A, response selection). These figures corresponded to 1% and 14% of items, respectively. It would be interesting to explore these differences in more detail to understand their causes. Clearly, figures as high as 14% are worrying as they suggest a potentially ineffective use of human resources (i.e., 14% marked by a senior marker *and* by two others).

It is also worth noting that certain questions on each paper posed substantially more problems than others. For example, on maths A, question 7 caused response selection markers the least problems (4% adjudication) while question 22a caused them the most (32%). Similarly, on English reading, least problems for clerical markers occurred for question 3 (1% adjudication) while most problems occurred for question 16 (11%). Detailed investigations into the reasons why such differences occurred may help to detect problems that can be overcome in future pilots and trials. One possible explanation that was mentioned by supervising markers was an element of confusion over the protocol for recording null responses.

The median total number of items adjudicated for expert markers ranged between 37 and 131. However, it should be realised that these corresponded to a much smaller number of items that received double marking. In fact, these medians corresponded to 10% (maths B) and 58% (English handwriting) of items, respectively. In effect, then, over one half of the double marked handwriting scripts were eventually triple marked (by a senior marker). Although we are discussing management data, this is, of course, a crucial index of

marking reliability. It suggests that there must be a large element of uncertainty concerning the (approximately) 90% of handwriting items that were not double marked.[7]

There needs to be a wider debate on the extent to which expert items will be double marked, by whom the double marking will be conducted, and the purpose that double marking will serve. If, for example, double marking were only to be used to provide monitoring data, then it might be decided that adjudication was actually inappropriate; a variety of alternative approaches could be adopted instead (e.g., arbitrarily awarding the first mark, awarding the mark of the more senior marker, or awarding an average of the two marks). For further discussion of this issue see 7.1.2.2.

5.1.1.5 Entire process

Question: ***how long did it take for the entire marking process to be completed?***

Tables 5.5.1A to 5.5.7A present data that encapsulate the entire Pilot, from first receipt of test papers (24 May 2001) until the end of the clerical marking period (15 June 2001). As indicated in 5.5.1A, it was not possible for NCS Pearson to mark all of the scripts that were eventually scanned within the time frame of the Pilot. The success rate was as low as one half for the expert items of the English writing test.

Tables 5.5 and 5.6 present summary data concerning the duration of the processing and marking stages of the Pilot. In addition to the total elapsed time, data are presented on the total amount of time during which script processing and clip marking took place. From these figures, notional values for the number of scripts processed, and clips marked, per minute were calculated. These are a notional reflection of the productivity of the Pilot given the available resources. It should, of course, be realised that the total active time is computed from the commencement of work each day to the end. The notional figure for productivity does not take account of the fact that resources (particularly human resources) were differentially available both within and between days.

---

[7] The data that were provided for English writing would have implied figures of greater than 100% being referred for adjudication. As such, they have been deemed erroneous and, therefore, unavailable. From what we know of English writing, however, there is little doubt that these figures would have shown a considerable degree of unreliability.

Table 5.5 The duration of processing.

	Begin	End	No. Scripts Processed	Total Active Time	Mean No. Scripts Per Minute
DWS Batch Builder	25 May	19 June	23,697	87 h 26 m	4.5
DWS Scan Master	29 May	19 June	20,490	181 h 16 m	1.9

Table 5.6 The duration of marking.

	Begin	End	No. Clips Processed	Total Active Time	Mean No. Clips Per Minute
Response selection	29 May	15 June	345,311	190 h 10 m	30.3
Clerical	29 May	15 June	332,198	188 h 6 m	29.4
Expert	29 May	11 June	114,120	158 h 36 m	12.0

The figure for mean number of clips per minute indicated, once more, that expert items took around three times longer to mark than clerical and response selection items (which took a similar amount of time to each other).

5.1.2 Processing accuracy

The second issue to consider was the extent to which procedures of the Pilot either increased or decreased the accuracy of processing of scripts through the system. This required the collation of evidence concerning error rates at various stages of the new technology marking process.

Unfortunately, very little information was forthcoming in relation to error logs. The only error log item provided concerned a script that had been damaged in transit back to the conventional marker (see Table 5.6.1A). This was resolved immediately by printing out a copy from the database. NCS Pearson noted that no further error logs could easily be provided.

Generally speaking, pupils' scripts appeared to be effectively processed and well accounted for, throughout the entire new technology marking process. It is assumed that this was facilitated, in particular, by the logging of script barcodes at all major stages.

5.2 Management data from conventional marking

Comparable data from conventional marking systems within the UK were hard to find. This was not simply for the obvious reason that the stages of conventional marking are different; it was also because detailed information on speed and accuracy by stage is typically not produced as a matter of course. As such, the discussion below addresses the comparative issue inadequately, but as well as was possible.

5.2.1 The appointment of markers

It was noted in Section 2 that NCS Pearson had problems recruiting sufficient numbers of appropriately experienced markers. However, it should also be noted that this was true for the Y7PTs more generally, as indicated within the conventional Recruitment and Retention reports, particularly for English.

5.2.2 Script receipt processing

The duration of conventional marking is heavily dependent upon the speed with which scripts can physically pass from one stage to the next. During new technology marking the physical script is far less central and the duration of the entire process is less at the mercy of transportation threats, such as postal delivery failure.

The new technology process invests time and effort at the beginning of the external marking process, to reap time and effort savings toward the end. Thus, we might expect a relatively slow pre-marking phase, due to the additional components of guillotining and scanning. For conventional marking, once a school has batched its scripts up, all that is required before markers begin work is for a delivery agency such as Parcelforce to deliver them.

Inevitably, the pre-marking phase of the Pilot was delayed to some extent, relative to the speed with which conventional marking was able to commence: new technology marking simply cannot begin until a sufficient number of scripts have been batched, guillotined and scanned. The results presented earlier gave a rough indication of turn-around times for the pre-marking phases, but it will not be possible to establish with any accuracy the speed with which scripts can pass through the batching, guillotining and scanning stages

until the system has been tested at a high capacity, with optimally designed scripts and without technology 'teething' problems.[8]

It is worth bearing in mind that much of the 'pre-marking' phase can actually run in tandem with marking. Thus, it would not be appropriate to set targets for subsequent new technology contractors to batch, guillotine and scan all, or even the majority, of scripts within an unduly narrow time frame. Instead, they need simply to demonstrate capacity for generating, within a short time frame (i.e., a small number of days), a sufficient *corpus* of clips to enable all markers to commence marking and to deliver a sufficient *flow* of clips to ensure that the pool is maintained at a level that supports all functions effectively.[9] The speed with which these procedures need to be undertaken will be a function of the marking demands to be made upon the system; that is, the larger the number of markers to begin marking, the faster the pre-marking phases will need to operate.

5.2.3 Marker standardisation

There was little post-training marker standardisation within the procedure of the Pilot.[10] As such, any comparison with the duration of the conventional marking process is problematic because considerable time is usually spent achieving this function. For the 2001 Y7PTs, conventional markers were to have received scripts from schools by 19 May (mathematics) and 23 May (English) and to have sent their first samples by 21, and 25 May, respectively. However, the deadline for team leaders to give feedback to markers on the first sample was not until almost a week later (27 and 31 May).

Clearly, there is scope for new technology procedures to accelerate this stage of the external marking process very significantly. The opportunity for partly, or entirely, automating the standardisation process is clear; but immediate access to a supervisor

---

[8] 'Optimally' is used to imply design changes that will be agreed between QCA and the new technology contractor (not necessarily those changes that will maximise the performance of the technology).

[9] Note that 'sufficient corpus' implies not only sufficient for markers to begin marking, but also sufficient to ensure that an appropriate number of monitoring items can be generated in a valid manner during the early stages of marking.

[10] That is, there was no formal standardisation beyond the S and T training scripts, although a small number of exemplar scripts were considered during the Pilot training session for expert markers.

within a centre is another way in which the postal service – and, hence, delay – can be avoided. As explained in Section 2, the question of how to manage the training, standardisation and supervision of on-line marking is crucial.

5.2.4 Script marking

In principle, the demands of on-line marking are less than the demands of conventional marking. In particular, there is no requirement to total marks, to transfer mark totals, to compute levels or to check marks and levels. Nor is there any requirement for completing or sending mark sheets. There are elapsed time gains in other ways, for instance, those related to the sorting, packaging and despatching of scripts and mark sheets. Therefore, we might expect the on-line marking stage to take significantly less time than the conventional marking stage, even assuming that exactly the same number of scripts would, ultimately, be marked.

The available data for comparing marking speed across systems were limited. A very rough notional number of minutes per script was presented in Table 5.2, with times for on-line marking ranging between 6.1 minutes (writing) and 20 seconds (arithmetic). Similar figures are not computed for conventional marking, nor are recommendations formally specified. However, the NFER English team (which developed the 2001 Y7PT) provided estimates for the time that a marker might be expected to spend on each script:

- reading – around 4 minutes;
- writing – just under 5 minutes;
- spelling/handwriting – just over 1 minute.

These figures are reassuring as they are clearly in the same ball-park as the earlier data from the Pilot. However, it should not be concluded that the on-line times (which were slightly slower) imply that on-line marking is not as fast. There are simply far too many uncontrolled variables to conclude anything other than that the figures are in the same ball-park and that this is reassuring.[11]

---

[11] In the 2000 marker evaluation report, markers were asked how long they spent undertaking all aspects of marking (including administration, marking, packaging, etc.) and how many scripts they were allocated. Dividing mean responses resulted in a figure of 19 minutes per script. Similarly, the Y7PT conventional marking panel of 20 June worked at a rate that equated to around 11 minutes per script for English and

5.2.5 Post-mark processing

While losing time during pre-marking phases – through the requirement for batching, guillotining and scanning – the procedures of new technology marking have the potential to make much of this back through efficient automation during post-marking phases. In particular, the automatic processing of marks (aggregation and level computation) means that this stage no longer makes 'real-time' demands. Furthermore, the fact that marks are directly entered means that there are no additional cost, resource, quality or time burdens of data input.

As mark sheet processing alone can take the best part of 3 to 4 weeks for national curriculum tests, the post-mark processing technology of on-line marking would clearly lead to significant time savings.

5.2.6 Processing accuracy

One of the more embarrassing problems facing national curriculum assessment is the loss of scripts or batches during the marking period. At the most extreme this has occurred when Parcelforce vans have been stolen. On other occasions entire batches have simply not been delivered, with no good reason provided. More commonly, individual scripts are lost between school and marker, or vice versa, with both parties insisting that the problem was not at their end.

While on-line marking will not eliminate this (at least until tests are completed on-line), it will inevitably help to overcome it by reducing the number of occasions that a script has to travel by post. It should also reduce the risk of loss, once a script has been scanned, as it will be possible to reproduce it from the electronic copy (as actually happened during the Pilot). Of course, this assumes that the electronic system is maintained effectively, for example, with regular back-up of servers.

One area in which there is still a threat of reduction in processing accuracy concerns the quality with which scripts are reproduced electronically. Instances outlined in Section 2 included writing responses having been presented in the wrong page order, or clips having been presented upside down, or clips simply being unreadable.

---

around 8 minutes per script for maths. Again, these are ball-park figures and cannot be used for direct comparison.

5.3 Analysis of the measurement data

Study 5 was focused primarily upon measurement data, where this was taken to be quantitative information on the accuracy of results arising from the marking of Y7PT scripts. The central question of study 5 was whether the implementation of new marking technology during the Pilot helped to eliminate errors or inconsistencies that inevitably creep in when scripts are marked by humans. These might relate to:

1. the evaluation of item responses by markers (i.e., errors due to not applying the mark scheme in the manner specified by the Chief Marker, or inconsistencies due to alternative judgements of the worth of item responses);

2. the addition of script marks by markers (e.g., incorrect totalling of marks on script or mark sheet, incorrect conversion of marks to levels, or incorrect scaling of spelling);

3. the completion of script and mark sheets by markers (e.g., failure to mark questions, correct marking which is recorded incorrectly, or incorrect transfer of marks to front of script, or incorrect transfer of marks from script to mark sheet);

4. the input of mark sheet data by clerical staff.

5.3.1 Methodology

In order to provide data for comparative purposes, the marks that had been awarded by the conventional markers were input by NCS Pearson data entry staff. These were input at the question level, which enabled some detailed statistical analyses to be carried out.[12] As marks awarded to responses were not visible on item clips, the on-line marks were awarded independently of the conventional ones. This enabled a more valid comparison of marking consistency between the conventional and the new technology systems.

Additional question level data were also produced, for each of the clerical and expert markers, when double marking had occurred. This made it possible to consider the consistency of marking within the new technology system. As explained earlier, the intention of study 5 was not simply to compare the quality of on-line marking with the quality of conventional marking, but also to illustrate the kinds of measurement data that

---

[12] As NCS Pearson employed a double-entry procedure, it was assumed that the final conventional data were relatively free of data input error. There was no empirical investigation of this proposition.

the new technology can routinely yield and to provide baseline information on the quality of on-line marking for future pilots and trials.

5.3.2 Inferential caveats

It is essential that the following analyses be contextualised by certain technical caveats. These should be taken into account before drawing wider inferences from the data.

5.3.2.1 The nature of the data

All of the analyses presented below are based upon databases that were provided for the NFER, by NCS Pearson, in accordance with the *Specification for Data Collection and Analysis*. Four sets of databases were requested:

1. six pupil-level databases, corresponding to each of the six papers (containing question-level mark data from both the conventional and the on-line process);

2. five marker-level databases, corresponding to the five papers with clerical questions (containing question-level between-marker reliability data – the number of clips per question for which the marker agreed with a re-marker – from the on-line process);

3. five marker-level databases, corresponding to the five papers with expert questions (containing question-level within-marker reliability data – the number of clips per question for which the marker provided the same mark a second time – from the on-line process);

4. five marker-level databases, corresponding to the five papers with expert questions (containing question-level between-marker reliability data – the number of clips per question for which the marker agreed with a re-marker – from the on-line process).

5.3.2.1.1 The marker-level databases

Initial checking of the fifteen marker-level databases uncovered an apparent anomaly. The Evaluation team assumed that there would be three possible data categories for each clerical marker on each question: the number of items for which the marker awarded the same mark as another; the number of items for which the marker awarded a different mark from another; the number of items for which the marker sent the clip directly to a supervisor without marking. In fact, data were not available for the last of these three

alternatives. NCS Pearson explained that, once marked by a supervisor, a clip is no longer attributed to the 'passing' marker but to the supervisor.[13]

While the basic 'agree versus disagree' data are interesting in their own right, it should be noted that they are importantly contextualised by the proportion of items for which a marker passes clips directly to a supervisor. If a marker considers only those clips for which she is certain of the mark, she might end up with highly reliable marking statistics, but from marking only a relatively small number of clips. Further, having marked only the easiest questions, her marking statistics would not be directly comparable with those from another, less conservative, marker.[14]

5.3.2.1.2 The pupil-level databases

One of the aims of the measurement analyses was to determine the extent to which conventional markers leave mark boxes empty, fail to total marks, etc.. This required the provision of 'raw' databases from NCS Pearson with missing or technically invalid data included. Inspection of these databases revealed blank and invalid data for each of the six

---

[13] Although NCS Pearson proposed that marks for clips 'passed' to supervisors were not attributed to those markers (but to their supervisors), a close scrutiny of Table 5.3.2A against Table 5.14.1A suggests a different story. The three markers of 5.14.1A were credited with having marked 2128 items (4256 clips) for each question. However, 5.3.2A notes that, for each question, a certain number of those 2128 clips had actually been sent to a supervisor. This implies either that the supervisor's mark was credited to the 'passing' marker and not to the supervisor, or that the mark was credited to both. Whichever, this practice would, in principle, lead to an over-estimate of marking quality for the 'passing' marker. Of course, given the small numbers of 'passed' clips in 5.3.2A, this over-estimate would be very slender. However, if higher numbers happened to be involved (if, for example, expert markers were given the opportunity to 'pass') then the problem might become significant. It is assumed that this kind of problem did not also affect data for adjudication. Incidentally, it is worth questioning whether such 'errors of attribution' might impact upon the payment to markers (if they were paid by item rather than by time).

[14] NCS Pearson explained that there was "an audit trail that includes where items have been sent to a supervisor but that was not used in deriving this data". In future years, if it proved useful, one option might be for supervisors to have access to statistics on clips sent to supervisors alongside those for marker reliability, to identify markers who were not necessarily marking to a high standard but were simply passing on a high proportion of items.

papers, although particularly for English reading. A summary of anomalous data from these databases is presented in Table 5.7A (Appendix 5.2).

NCS Pearson had specified to data entry staff that missing mark data be recorded with a '\'; yet it appeared that '/' was also accepted (as these were also present on the databases). From 5.7A, it is clear that reading had by far the largest number of pupils for whom there was at least one blank entry in the mark fields. The vast majority of these were recorded with a '\', although '/' was also recorded and some cells were simply blank. The reason for the particularly high rate of blanks for reading was not apparent.

For the purpose of analysis, it would have been safe to assume that any data that had been entered as '\' genuinely represented an instance of a marker not having recorded a mark.[15] However, it would not necessarily have been safe to assume that these blanks genuinely represented zero marks, as markers might simply have forgotten to record the mark that they considered the responses to be worthy of, or might even have forgotten to mark those questions. As such, it was decided to eliminate from the main analyses any pupil for whom there was at least one blank or invalid entry in any of the conventional or new technology fields. This was only a serious problem for reading, as the sample was reduced to 37% of the original number of pupils; for the remaining papers, at least 89% of the original number of pupils were included in the main analyses.

Due to this elimination, those pupils whose data were analysed for reading might, conceivably, have been of higher ability than those whose data were not analysed (assuming that many of the blanks did indeed correspond to instances of pupils not providing responses to questions). This might also be true for the other papers, although to a lesser extent. Of course, data from precisely the same pupils were analysed for both conventional and new technology marking, so comparative analyses were not affected.

5.3.2.1.3 The nature of the samples

Finally, it should be noted that the Y7PTs are intended only to sample lower achieving students (who fail to achieve level 4 in the previous year's key stage 2 tests). This had two main implications for the analysis of results from the Pilot:

---

[15] Recall that conventional marks were double entered and the '\' would, presumably, have been entered twice.

1. The mark distributions, for both conventional and new technology marking, were likely to be of restricted range, in comparison with the corresponding key stage 2 test mark distributions. This would be likely artificially to attenuate any observed correlation coefficients, at least to a small extent. In fact, for both maths and English, a comparison of mean marks between the 2001 Pilot sample and the 2001 key stage 2 cohort (who took the same test) revealed considerably lower marks for the Pilot sample, thereby confirming that they were generally of lower ability.

2. On average, pupil responses to each question of each paper were likely to be less sophisticated than corresponding key stage 2 test responses. This would be likely to make them easier to mark accurately (especially with the relative over-representation of null responses that might be expected from this sample). Conceivably, then, this might have the opposite effect of artificially inflating the observed correlation coefficients between new technology and conventional marking, at least to a small extent. It might also mean that baseline data on marker agreement within the on-line system might not extrapolate directly to a scale-up in which the full cohort was tested.

5.3.2.2 The validity of the data

Having clarified the nature of the data and samples that provided the basis for study 5, it is also important to note further issues that might have impacted upon the validity of marks awarded through the conventional and new technology processes.

5.3.2.2.1 The validity of conventional marks

The main point to note concerning the quality of conventional marking is that scripts had not been borderlined before marks were input by NCS Pearson. As borderlining results in changes to the marks of a significant number of pupils, the profile of conventional marks used in the Pilot was somewhat different (at least for certain pupils) from the profile that would ultimately have resulted for the Pilot sample. Similarly, errors on scripts are picked up when the mark sheet is completed and final clerical checks are carried out. To the extent that such checks may not have been undertaken for the Pilot, the conventional data may have somewhat under-represented the general quality of conventional marking.

Of course, these borderline and checking stages by no means eliminate conventional marking errors, as the incidence of successful review requests demonstrates.

5.3.2.2.2 The validity of new technology marks

A second question is whether the quality of data arising from the on-line marking was a fair representation of the quality of marking that might be expected in future new technology pilots, trials or roll-outs. There were two main issues that bore on this concern.

First, the markers used for on-line marking were less experienced than the typical national curriculum marker. This is likely to have had some impact upon the results obtained. Unfortunately, it was not possible to quantify the extent to which this might have occurred.

Second, there was a problem with elements of pupils' responses to questions occasionally being recorded beyond the clip image areas (see Section 6). These elements were simply not available to markers for scrutiny. To the extent that pupils' responses may have received different marks, had full-page images been accessible to markers, the validity of the new technology marks will have been compromised. At least in the case of English, the policy adopted in relation to marking such items was one of 'if it ain't there, don't mark it'.

5.3.3 Between-system comparisons

The principal comparison of results between conventional marking and on-line marking concerned final marks. Agreement between subject, paper and sub-section mark totals from the two systems was investigated using correlation and mean mark difference statistics.[16] The extent of agreement between marks awarded for each item of each script was subsequently analysed using concordance statistics, and agreement between levels awarded for each of the subjects overall was also considered.

It is important to realise that the following three sets of analyses simply indicate the extent of agreement between conventional and on-line marking. They do not indicate whether one was more or less accurate than the other (nor, indeed, whether their accuracy differed at all). This is because there was no independent external reference point determining the 'correct' mark for each script or item.

---

[16] Note that, whenever a script total was analysed for conventional marking, it was computed from the sum of the item marks and not taken from the total mark recorded on the script by each marker (as these may be error prone – see 5.3.5 and Table 5.12A).

5.3.3.1 Correlation

Coefficients of correlation essentially indicate the degree of consistency of ranking between two sets of data for the same cases. For the Pilot, the 'cases' in question were individual pupils and the two 'sets of data' were the marks awarded to them during the conventional and new technology processes, respectively. These were considered for each of the papers separately (as well as for the subject overall) and also for sub-sections of each paper relating to response selection, clerical and expert markers.

The full correlation statistics are presented in Table 5.8A (Appendix 5.2) and these are summarised in Table 5.7. The figures in brackets within each cell indicate the maximum mark available for each subject, paper or sub-section.

Table 5.7 The correlation between conventional and new technology marks.

	Subject	Script	Expert Questions	Clerical Questions	Response Questions
Reading	**0.90** (100 marks)	**0.94** (50 marks)	**0.91** (38 marks)	**0.96** (11 marks)	**0.98** (1 mark)
Writing		**0.62** (35 marks)	**0.62** (35 marks)		
Spell./Hand.		**0.93** (15 marks)	**0.56** (5 marks)	**0.97** (10 marks)	
Maths A	**0.99** (100 marks)	**0.98** (40 marks)	**0.95** (11 marks)	**0.98** (13 marks)	**0.98** (16 marks)
Maths B		**0.98** (40 marks)	**0.97** (16 marks)	**0.98** (15 marks)	**0.96** (9 marks)
Maths Arith.		**0.98** (20 marks)		**0.98** (5 marks)	**0.98** (15 marks)

5.3.3.1.1 Evaluation of English coefficients

There is not a great deal of data from conventional marking systems in the UK against which to compare the analyses presented above. In his studies of marking reliability for the GCE A-level, Murphy (1978) recorded the following paper-level correlation coefficients for the three AEB English language papers of 1976: 0.73, 0.75 and 0.76. In a subsequent paper (Murphy, 1982), he considered the marking of AEB English language O-level across a period of three years. He recorded marking reliability coefficients of 0.75 (paper 1) and 0.91 (paper 2) for 1976, against 0.76 (paper 1) and 0.93 (paper 2) for 1979. Newton (1996) presented more recent data for GCSE English from the same examining board. The two main English papers for the 1994 examination (one per tier) were marked according to 4 criteria: reading 1 (short answer questions); reading 2 (marks

awarded for impression of 'reading' ability demonstrated in extended responses); writing 1 (marks awarded for impression of 'writing' ability demonstrated in extended responses); and writing 2 (marks awarded for presentation of extended responses). The corresponding marking reliability coefficients ranged between: 0.85 and 0.91 (r1); 0.70 and 0.82 (r2); 0.74 and 0.92 (w1); 0.80 and 0.89 (w2).

More recently still, and of more direct relevance, Whetton, Twist and Savage (1998) presented results of a project investigating various mark scheme models for KS2 English writing. The estimated marking reliability coefficients for the four models were as follows: 0.78 for the conventional model; 0.75 for the holistic model; 0.83 for the paired holistic model; and 0.81 for the grid model. The conventional model was essentially similar to the current mark scheme.

All of the above coefficients relate to English papers, or sub-sections of papers. Generally speaking (for technical reasons related to the process of aggregation) subject-level coefficients are slightly higher.

The coefficients observed for reading during the Pilot were encouraging as they exceeded even the best of those presented above. The figures for response selection and clerical markers were particularly high, as might be expected considering that these questions were selected on the basis of simplicity to mark. Yet even the figure of 0.91 for expert markers alone was very good. At 0.97, the coefficient for spelling was also very encouraging (although it must be admitted that no spelling reliability data from conventional marking alone were available for direct comparison). The coefficient for writing was less encouraging, being substantially lower than even the worst of the comparable figures presented above. Finally, the coefficient for handwriting was disappointing (although, again, no handwriting reliability data from conventional marking alone were available for direct comparison).

The high correlation coefficients for reading provide evidence that mutually validates the marking from both systems. For writing, with lower coefficients, the picture is less clear. It could be that the conventional marking was reliable, but the new technology marking less so; alternatively, the converse could be true; indeed, both systems could have been performing sub-optimally.

Even if it could have been demonstrated that the lower coefficients for writing were entirely due to the quality of marking arising from the new technology system, this would not necessarily have implied that the fault was with the system, per se. Recall that the

Pilot markers were not as experienced as ideally they should have been. This, alone, could provide an alternative explanation for any marking reliability problems.[17]

5.3.3.1.1 Evaluation of maths coefficients

Murphy (1982) also included the details of an analysis of the 1977 examinations in O-level mathematics and A-level pure mathematics. For both of these subjects the marking reliability correlation coefficients were very high. Two of the three O-level papers had a coefficient of 1.00 (although one of these was a computer-marked objective test), and the other had a coefficient of 0.99. One of the three A-level papers had a coefficient of 1.00, another had a coefficient of 0.99, and the third a coefficient of 0.98. Clearly, the standards of reliability were very high for maths; in fact, maths was the most reliably marked of all subjects. Newton (1996) found very similar results for the six main GCSE maths papers (two per tier) of 1994 – none of the correlation coefficients fell below 0.99.

The correlation coefficients for the maths papers of the Pilot were nearly as high: the paper level coefficient for each of the three maths papers was 0.98. This would seem to provide evidence supporting a mutual validation of conventional and new technology marking. Note that even the difference between expert and clerical/response selection markers was not that great for maths.

5.3.3.2 Mean mark difference

It is conceivable that the task demands of conventional or new technology marking may have led to a consistent bias in the marks awarded to scripts (i.e., a general lenience or general harshness). This possibility would not be apparent from correlation coefficients alone, which is why mark differences needed also to be investigated.

---

[17] Instead of randomly sampling clips for expert double marking from the entire expert item pool, it would have been possible to sample all expert items from a random sample of pupils. If this had been done (recalling that all clerical and response selection items were double marked) it would have been possible to compute marking reliability coefficients for the new technology process in isolation. These could then have been compared 'directly' with those arising from conventional marking reliability studies. It will be important to consider this kind of investigation in future pilots or trials (remembering that the quality of markers must also be controlled for the between-system comparisons to be meaningful).

Table 5.8 The mean mark difference between conventional and new technology marks (an '*' denotes insignificant at $p < 0.05$).

	Subject	Script	Expert Questions	Clerical Questions	Response Questions
Reading		**2.11** (50 marks)	**2.04** (38 marks)	**0.06** (11 marks)	**0.01** (1 mark)
Writing	**0.90** (100 marks)	**0.25 *** (35 marks)	**0.25 *** (35 marks)		
Spell./Hand.		**0.31** (15 marks)	**0.17** (5 marks)	**0.14** (10 marks)	
Maths A		**0.22** (40 marks)	**0.01 *** (11 marks)	**0.10** (13 marks)	**0.10** (16 marks)
Maths B	**0.99** (100 marks)	**0.27** (40 marks)	**0.12** (16 marks)	**0.04** (15 marks)	**0.12** (9 marks)
Maths Arith.		**0.12** (20 marks)		**0.00 *** (5 marks)	**0.12** (15 marks)

The possibility of general lenience or harshness was evaluated using 2-way repeated measure (same pupil) t-tests of conventional versus new technology marks. The results of these analyses are presented in full in Table 5.9A and summarised, above, in Table 5.8.

5.3.3.2.1 The differences across all comparisons

The mean mark differences were particularly intriguing, the first reason for this being that they all favoured conventional marks. That is, when new technology marks were subtracted from conventional marks, the mean of the differences was always positive. This was not only true at the paper level, but for all three different types of marker (even within the four cells where the mean mark difference was not statistically significant).

In the absence of a plausible counter-explanation for this unexpected result, the conclusion would have to be that the process of new technology marking results in the award of marginally lower marks than the process of conventional marking. This suggests that within one (or both) of the systems an unidentified demand characteristic leads to a particular bias in the marks awarded to pupils.

It is assumed that the unexpected result cannot be explained by a simple technical problem affecting the data provided by NCS Pearson. These analyses were conducted on databases restricted to only those pupils with valid data in all fields, so the anomalous entries mentioned earlier are not the explanation (nor is there reason to believe that restricting the databases to only those pupils with valid responses was responsible for the effect). One possible explanation might be that the new technology software (presumably

during the data manipulation, storage or output stages) introduced occasional errors into the new technology data files whereby, for example, item marks greater than zero were transformed into zeros. However, it is not clear why this might have happened, nor is there any evidence to think that it might have.[18] Nor is there any reason to think that the effect was due to any unanticipated characteristics of the markers who marked the scripts conventionally.[19]

Unfortunately, assuming that the effect was attributable to particular demand characteristics of either the new technology or the conventional marking system, it is not at all obvious what the explanation is. One possibility might be that conventional markers – seeing personalised scripts rather than de-personalised question responses – were more disposed to give the benefit of the doubt to candidates.

The finding is very intriguing and would have significant implications for a wider roll-out of the new technology system. It clearly demands further research.

5.3.3.2.2 The difference for English reading

The second reason why the results were intriguing concerned English reading specifically. Not only was there a significant mark difference, but this was an average difference of 2 marks. If two conventional markers had marked the same sample of scripts, and their mean difference between marks awarded had been computed, then a 2 mark difference would not necessarily be surprising. It would simply mean that one marker was more lenient than the other. When extrapolated to the Pilot, though, the implication would have to be that the group of conventional markers was more lenient than the group of new technology markers. As the numbers of markers in each group was not small, this is a far more surprising outcome.

Moreover, when the difference between conventional and new technology reading results was analysed at the question level, the mean mark difference in favour of the conventional system was replicated across 29 of the 31 questions. This further demonstrated the

---

[18] Earlier, the possibility of expert on-line markers accidentally 'Committing' the default entry of 'no response' was raised. While this is still a possibility, it would not explain mean mark differences for clerical or response selection markers (for whom double marking would have identified such errors).

[19] Incidentally, as borderlining tends to inflate the overall mark profile, this effect may have been even more pronounced had borderlining been implemented by the conventional markers.

consistency of effect for both expert and clerical markers (the one response selection question also favoured conventional). This is despite the fact that non-expert items were double entered.

This is a highly irregular finding and there is no clear explanation for it. Once again, and more emphatically so, there is a need for further research to establish whether this effect is replicable. If it were to be replicated, then there would be significant implications for a wider roll-out of the new technology system.

5.3.3.3 Question-level concordance

To provide a more detailed indication of the nature of agreement between conventional and new technology markers, an index of concordance was computed for each question. This was based upon a statistic called Cohen's kappa. Also computed were agreement statistics for each question which represented the percentage of items for which the new technology mark was in precise agreement with the conventional mark.

Table 5.9 summarises data presented in Tables 5.10.1A to 5.10.6A. It records the mean percentage of items for which there was an exact agreement between conventional and new technology markers (averaged across questions in each group). These figures are presented separately for questions with different maximum marks, as there is more chance of disagreement on a multiple mark scale than on a single mark scale.

Table 5.9 Mean percentage exact agreement figures (the number of questions from which each average was computed is in brackets).

	Resp.	Clerical			Expert			
	1	1	2	3	1	2	3	>3
Reading	**99** (1 qn)	**99** (6 qns)	**96** (1 qn)	**94** (1 qn)	**87** (9 qns)	**77** (10 qns)	**78** (3 qns)	
Writing								**54** (3 'qns')
Spell./Hand.		**98** (20 qns)						**54** (1 qn)
Maths A	**99** (16 qns)	**98** (9 qns)	**98** (2 qns)		**96** (3 qns)	**96** (4 qns)		
Maths B	**98** (9 qns)	**99** (15 qns)			**97** (4 qns)	**96** (6 qns)		
Arithmetic	**99** (15 qns)	**99** (5 qns)						

The data complement and extend the correlation coefficients presented earlier. Once again, the figures for mathematics questions were very high. The average percentage

agreement across the clerical and response selection questions did not fall below 98%. Even for the expert questions, the lowest average percentage agreement was 96%.

For English reading, the figures for response selection and clerical questions were similarly impressive, with the percentage agreement for even the three-mark clerical question reaching 94%. The percentage agreement figures also more clearly indicated that the problems of consistency for reading were more serious for expert markers than for clerical and response selection markers (as might be expected). For the single mark expert questions the average percentage exact agreement figure was 87% and the average percentage disagreement was larger still for the two and three mark questions.

For both writing and handwriting the average percentage exact agreement was less impressive, at only 54%. Once again, though, this does not necessarily imply that the conventional marking was less consistent than the new technology marking, or vice versa; it simply means that there were problems of marking consistency for one or both of the systems.

Little more will made of the data within Tables 5.10.1A to 5.10.6A within the present report. However, it is likely that the question-level information will be of particular interest to senior markers and test development teams. By delving more deeply into the characteristics of questions with poor concordance it may be possible to identify patterns of marking error that can feed into future developmental work.

5.3.3.4 Subject-level concordance

Finally, to provide a stark overall comparison of conventional and new technology marking, the absolute agreement between levels arising from the two approaches was examined.

Table 5.11A reveals that 5% of students would have received different levels had the new technology marks been used for maths and 27% of students would have received a different level for English. This high figure for English primarily relates to the large mean mark difference for reading. In fact, 20% of students would have received a lower level for English, had the new technology marks been used instead of the conventional marks, while only 7% would have received a higher level.

5.3.4 Within-system comparisons

In addition to analyses comparing the quality of marking between systems, it was possible also to conduct analyses to compare the quality of marking within the new technology

system (i.e., between the on-line markers). The primary intention of these analyses was to provide baseline information for future pilots and trials.

Three sets of databases were collected for this purpose:

1. between-marker reliability databases for clerical markers (relating to all items marked by each marker);

2. within-marker reliability databases for expert markers (relating to a subset of around 3% to 4% of each marker's allocation that they re-marked for monitoring purposes);

3. between-marker reliability databases for expert markers (relating to a subset of around 20% of each marker's allocation that was re-marked by other markers for monitoring purposes).

5.3.4.1 Clerical markers: between-marker reliability

Tables 5.14.1A to 5.14.6A (Appendix 5.3) present data concerning the performance of individual clerical markers. For each paper, these include: the numbers of clips marked; the numbers of questions for which clips were marked; and the extent of agreement, across items marked for each question, between each marker and the various re-markers.

As explained fully in Table 5.14.3A, clerical markers were assigned to question groups and, within each group, each marker marked all responses from a particular pupil to the questions within that group. Thus, Marker 1 awarded marks for 9 questions from maths A, having been assigned to two question groups (with 5 and 4 questions, respectively).

Of most significance are the data in the maximum, minimum and mean percentage of 'clips in agree column' columns. These refer to the percentage of items per question for which each marker agreed exactly with the re-marking markers.[20] The data are presented in full to form a baseline for future marking pilots and trials.

---

[20] Note that exact agreement was computed for all of the following analyses, despite the fact that different questions had different maximum marks (and, therefore, different chance probabilities for disagreement). As such the data may not be considered precisely comparable across questions or across markers. Indeed, to the extent that certain markers may have been assigned to harder question groups than others (that is, harder to mark consistently) then any comparisons between markers should always be made cautiously. This is another issue that future contractors will have to consider when developing marker reliability statistics for monitoring purposes.

These data are actually very similar to the agreement data of 5.3.3.3 although, of course, the difference is that they are broken down by marker. This provides an indication of the difference in marking quality between the markers. In fact, the range in average marking accuracy across the clerical markers was not great during the Pilot. Tables 5.13.1A to 5.13.4A reproduce only the mean percentage of 'clips in agree column' columns, for all papers and each marker type. These data are further summarised in Table 5.10 below.

Table 5.10 The median percentage exact agreement across markers (median across markers of mean across questions).

	Clerical Between	Expert Within	Expert Between
Reading	97.3	90.2	88.5
Writing	-	66.7	58.3
Spell. & Hand.	97.7	57.8	41.4
Maths A	95.9	96.8	95.5
Maths B	97.2	97.1	96.3
Arithmetic	98.1	-	-

As is clear from Table 5.10, the typical level of agreement between clerical markers ranged from 96% to 98% across papers. The lowest mean percentage agreement for any individual English clerical marker was 97% (reading) and, for any maths clerical marker, was 94% (maths A).

5.3.4.2 Expert markers: within- and between-marker reliability

Single (rather than double) marking was employed for expert questions and re-marking only occurred for the purpose of monitoring. As such, data for expert markers (presented in Tables 5.15.1A to 5.16.5A) also included information on the number of items for which monitoring was undertaken for each marker.

This information is interesting in its own right because, even for the same marker, there was considerable variation between questions in the proportion of items sampled for re-marking. For example, it was common for certain questions within a marker's allocation not to be monitored for within-marker reliability.[21] However, even for between-marker

[21] See 5.15.1A to 5.15.5A column 'No. questions with clips marked but not re-marked'. This was not true for between-marker reliability monitoring. It should be noted that within-marker reliability was more of an 'added extra' to the system specification and not designed with the same intent as the between-marker reliability function.

monitoring, there was considerable variation across questions in the proportion of items re-marked. In an exactly similar way there was considerable variation between markers in the proportions of items re-marked. For example, one English reading marker had 46% of items re-marked for between-marker reliability, while another had only 14%. Similarly, one maths A marker had 62% of items re-marked, while another had only 16%.

Once again, the most important data were contained in the maximum, minimum and mean percentage of 'clips in agree column' columns (as summarised in Table 5.13.2A, 5.13.4A and in Table 5.10 above).

As noted earlier, writing and handwriting exhibited the lowest levels of agreement between conventional and new technology markers. In the same way, writing and handwriting exhibited the lowest levels of agreement within the new technology system, both for within-marker and between-marker comparisons. The levels of percentage exact agreement were particularly poor for handwriting, where the median between-marker agreement was only 41% (58% for within-marker agreement). For writing, the comparable figure was 58% (67%). The fact that the between-marker figure was so low for handwriting (even in comparison with levels of agreement between conventional and new technology markers) suggests that this was a particular problem for the on-line markers. Of course, this should not necessarily be attributed to a characteristic of the system, as it might simply be that the sample of markers employed were not particularly good at marking handwriting. Further research will be required to explore possible explanations for the problems encountered.

From Tables 5.13.2A and 5.13.4A it is clear that levels of agreement were generally higher for within-marker, rather than between-marker, reliability. This is precisely as would be expected from any marking system. Also apparent from these tables is the fact that levels of agreement differed across markers; that is, some markers were more reliable than others. Handwriting is a good example of this, with mean levels of agreement ranging from 0% to 62% for between-marker comparisons and from 20% to 87% for within-marker comparisons.[22]

---

[22] The figure of 0% came from Marker 51, who marked 52 handwriting clips with 12 re-marked for between-marker monitoring. Although this marker appeared to have a within-marker agreement level of 50%, this arose from only 2 re-marked items (and so is questionable).

For writing, the range in levels of mean percentage exact agreement across markers was similar (around 60% to 70%), but arose from a higher baseline. Thus, mean levels of agreement ranged from 30% to 92% for between-marker monitoring and from 33% to 100% for within-marker monitoring. The range was much smaller for reading, from 74% to 100% for between-marker monitoring and from 76% to 98% for within-marker monitoring.

The comparison of figures for within- and between-marker agreement is a potentially interesting source of information. Marker 102 (Table 5.13.2A) is a case in point. According to the within-marker statistic, s/he was the 3rd most consistent handwriting marker whereas, according to the between-marker statistic, s/he was the 3rd least consistent marker. This suggests that Marker 102 was marking either consistently more harshly, or consistently more leniently, than the other markers.

The identification of particularly low levels of agreement, suggests that certain of the on-line markers may have been marking sub-optimally (e.g., Marker 51 for handwriting, or Marker 23 for reading). With the immediate production and analysis of monitoring data, such markers can be detected before they mark too many items incorrectly. This is a very important potential strength of the new technology system; its realisation will depend on a sufficient amount of the right kind of data being produced at appropriate times and used in an appropriate manner by the supervising markers.

As would be expected from earlier results, the analyses of monitoring data for the maths expert markers uncovered far fewer concerns. The lowest levels of mean percentage exact agreement fell no lower than 88% for any marker, while the medians (of these means) ranged from 96% to 97%. No maths markers appeared from the monitoring data to stand out as having had particular problems.

5.3.4.3 Absolute marking standards

The Chief Marker for maths, also an employee of NCS Pearson, grappled with the thorny issue of how to determine the extent to which markers during the trial were marking at an appropriate standard. His approach was to work backwards from the conventional criteria for establishing marking quality, considering Absolute Mark Difference (AMD) tolerances for the first sample.

When assessing each conventional marker's first sample, the supervising marker records the AMD between the mark that that they would have awarded to each answer on a particular script and the mark that the marker actually awarded. For each of the ten pupils

in the first sample, the sum of these question AMDs are calculated and then the Total AMD across all scripts is calculated from the sum of pupil totals. For the 2001 Y7PT in maths, to be classified in band A, a marker would have had to have had a Total AMD below 10; from 10 to 19 she would have been classified in band B; and for 20 or more she would have been classified in band C.

Of course, these conventional AMDs are computed from all questions on a paper whereas expert on-line markers are restricted to only a sub-sample; moreover, this sub-sample is restricted to those questions that are hardest to mark. The approach taken by the Chief Marker to overcome this problem was as follows:

1. for each e-marker, calculate an 'error rate' from the new technology monitoring data (T_{AMD}/S_M);[23]

2. for each marker, multiply their 'error rate' by ten times the sum of maximum marks available for expert questions within the paper under scrutiny (to arrive at an estimate of how large each e-marker's Total AMD-for-expert-questions (AMD_E) would have been had they been sampled conventionally);

3. assuming that the conventional marker ratio of AMD_E:AMD_{NE} is 6:3, project from AMD_E a value for AMD_{NE};[24]

4. for each e-marker, add these two AMDs to arrive at an estimated Total e-AMD;

5. directly compare these e-AMDs with the conventional thresholds (i.e., 9 for grade A).

According to this method, 5 of the maths markers were classified as grade A, 3 as grade B and 2 grade C.

The method appeared to be a useful approach to evaluating the quality of the on-line markers according to the standards laid down for conventional markers. It could be an

---

[23] T_{AMD} = total absolute mark difference, between marker and re-mark markers, across all items evaluated for between-marker reliability; S_M = sum of maximum marks available for each question, across all items evaluated for between-marker reliability.

[24] AMD_E:AMD_{NE} represents the ratio of AMD arising from expert questions to AMD arising from non-expert questions. This was established empirically, for conventional markers, through an analysis of 149 S1 sheets from the 2001 maths progress test.

effective starting point for considering the evaluation of marking quality for future pilots and trials.

5.3.5 Accuracy of conventional clerical tasks, checking and data input

A final analysis of accuracy – that was made possible by the input of conventional marks by NCS Pearson – concerned the accuracy of script annotation during conventional marking. Table 5.7A indicates that, for each of the six papers, there were a number of scripts that had no mark total recorded by the conventional marker (despite there having been marks to total). The percentage of scripts with this oversight ranged from 0.19% (mental arithmetic) to 1.09% (maths B).

Conventional markers are also prone to making addition mistakes. Table 5.12A records the prevalence of such errors by paper. They were least common for writing, where 0.01% of scripts had addition errors. It is not surprising that there were few addition errors for writing as there were only three marks to add. Errors were most common for reading (4.6%), maths A (2.6%) and maths B (2.4%), which all had substantially more marks to total for each script.

These figures are high. During 2000, levels were changed following a KS2 English R1 (clerical) review request for only 0.16% of the cohort. This may mean that the data presented above are not representative of the quality of conventional script addition for national curriculum tests more generally. This is a possibility, particularly as the progress tests do not have the same 'high stakes' attached. An alternative explanation might be that many of these errors are generally detected by markers, during final clerical checks, before scripts are returned to schools. A final possible explanation is that the frequency of script addition error is significantly under-represented by the national curriculum review data (i.e., that mark addition errors frequently go unreported – particularly when they favour, rather than penalise, pupils).[25]

In fact, given the data presented in Table 5.11, all of these explanations seem likely. Table 5.11 presents the results of an analysis of mark sheet errors which was computed by the NFER using data provided by the NDCA. The data are at the subject level and present the frequency of Error 1 and Error 2 (where the former reflects mistakes that did not impact upon a pupil's level and the latter reflects mistakes that did). An overall figure of

---

[25] Of course, a small amount of that 'under-representation' will be due to R1 clerical errors that occur in scripts submitted for R2 review.

2.41% for key stage 2 English is somewhat lower than the 4.6% recorded for the reading paper from the 2001 progress tests. Yet, it seems a more realistic representation than the 0.16% of pupils with supported R1 review requests.

Table 5.11 The number of pupils whose marks were affected by mark sheet errors during 2000.

		Error 1	Error 2	Any Error	No. participating
KS2 E	number of pupils	10926	4068	14994	623039
	% of all participating	1.75	0.65	2.41	
KS2 M	number of pupils	2061	1736	3797	623275
	% of all participating	0.33	0.28	0.61	
KS2 S	number of pupils	1798	1439	3237	622954
	% of all participating	0.29	0.23	0.52	
KS3 E	number of pupils	1970	2367	4337	579612
	% of all participating	0.34	0.41	0.75	
KS3 M	number of pupils	1780	1669	3449	581891
	% of all participating	0.31	0.29	0.59	
KS3 S	number of pupils	2018	2656	4674	580117
	% of all participating	0.35	0.46	0.81	

Presumably, given an appropriate aggregation algorithm accurately used, errors such as these would simply not occur at all during new technology marking. This would seem to be the baseline against which conventional marking should be compared.

5.3.6 Supplementary question analyses

Following a request from QCA, the following question analyses were conducted for each question from each of the six papers:

1. the computation of a simple facility index (mean mark divided by maximum);

2. the computation of a simple discrimination index (part-with-whole correlation).

The analyses were based purely upon conventional marks (only for pupils with valid, non-blank data across all fields). As the sample was restricted to the less able of the Year 7 cohort, the data would not be comparable with those that would have resulted if the entire cohort had been tested. The results are presented in Appendix 5.4.

The supplementary question analyses were not a formal component of the Evaluation contract and will not be discussed further in the present report.

5.4 Reflections

The goals outlined for Section 5 were: to illustrate the range of management and measurement data that it is possible to produce using the new marking technology; to provide a corpus of management and measurement data from the Pilot to use as a baseline for future pilots and trials; and to compare data arising from the Pilot with data arising from conventional marking systems.

5.4.1 Data, data everywhere...

In collaboration with NCS Pearson, it was possible to gather a large amount of management and measurement data from a variety of components of the Pilot system. In contrast to the lack of quantitative information from conventional systems (that have been in existence for many years and decades) a very clear conclusion emerged from the Pilot: new technology marking has great potential to enhance the effectiveness and defensibility of external marking through the production and analysis of large amounts of data on the conduct of its procedures. These data may be processed during the marking period, to determine whether progress is continuing as intended. Alternatively, they may be processed at the end of the marking period, as a record of the success of the operation. Perhaps the most promising advance, though, is their potential for supporting the work of managers, test developers, senior markers and researchers in determining how the system might be made more effective in successive years. The lack of detailed question- and marker-level data has restricted the potential for this kind of analysis in the past.

Ironically, though, the very capacity to produce large amounts of data may pose new risks for the external marking system, in particular:

1. the risk that data overload will burden and distract key players;
2. the risk that data will be mis-analysed, misunderstood or misused;
3. the risk that too much data at too low a level, or an inability to utilise data effectively, will cause senior markers to lose confidence in procedures associated with on-line marking.

These risks will probably be most acute when the new technology is required to accommodate to the idiosyncrasies of the UK external marking system. The use of data for monitoring expert markers is a prime example of this. There is a vast amount of monitoring information that could, in principle, be produced. But this does not mean that all of it ought to be produced as some data will be more useful than others; for example,

absolute mean difference figures will support inferences that non-absolute mean difference figures may conceal. Similarly, it will be important to establish when primary data sources need to be supplemented by additional contextual information; for example, when percentage agreement data need to be supplemented by information on the percentage of items 'passed' to a supervisor. Furthermore, it will be important when developing specifications for new statistical information, to ensure that data produced are fit for their intended purposes and valid in all circumstances (or that when potentially invalid data are produced they are flagged as such); for example, percentage agreement statistics that are computed from only a small number of re-mark incidents will need to be identified as such, or not produced at all.

The development, implementation and maintenance of data production systems and procedures will require input not simply from managers, database developers and senior markers, but also from those with expertise in the generation and analysis of psychometric data. Furthermore, once systems and procedures have been developed, it will be crucial that those who will use the data are given sufficient training in how to interpret and apply them appropriately. Data do not speak for themselves! New technology offers potential for the external marking system to embrace an extremely information-rich environment. It will be essential to ensure that the system facilitates, rather than obfuscates, the translation of data into information.

5.4.2 Messages from the management data

It will not be possible to determine definitively whether new technology has the capacity to enhance the processing speed and accuracy of the external marking system until it is tested:

- with a high volume of scripts, delivered on time;
- with markers that are demonstrably of the conventional standard, but also familiar with new technology marking;
- with procedures that embrace the full range of conventional processes;
- with procedures that exploit the full potential of the new technology;
- having overcome technological teething problems;
- with pupils sampled from across the full ability range; and
- with results provided for every single pupil in the sample.

However, evidence from management data arising from the Pilot did not give sufficient reason to doubt that the following speed and accuracy benefits might well accrue.

First, the management data did not provide compelling reason to doubt that time will be saved as:

- teachers will not have to sort completed scripts into order (although they may spend more time distributing personalised papers to pupils prior to each test);
- the postal system will be used less frequently;
- the pool of markers can be increased as non-experts are drafted in to mark non-expert questions (although the higher the proportion of double marking for experts, the smaller the potential time gain would be);
- markers will not have to add marks, compute levels, complete mark totals, fill in mark sheets, etc. (and checks on these procedures will be eliminated);
- data will be input directly by markers and not subsequently by inputters;
- assuming continued full-page image scanning, any review procedure could be initiated by telephone as the script would not need to be sent by post.

One of the major potential threats to the production of results in a timely fashion is the necessity for additional stages before the on-line marking commences; in particular, the batching, guillotining and scanning stages. Additional scanning 'exceptions' (which reached a maximum of 5.1% during the trial) only exacerbate this threat.

Yet the threat of these stages constituting a bottleneck should not be over-stated. First, they do not have to be entirely completed before marking can begin; it simply needs to be ensured that a sufficient corpus of clips is scanned and that a sufficient flow of clips into the system is maintained. The exact size of this corpus will largely be a function of the number of staff and machines devoted to batching, guillotining and scanning and the number of markers due to commence marking. This ratio will need to be established empirically. Second, even conventional markers do not begin the marking process proper until they have received confirmation of the acceptability of their first sample. To the extent that comparable 'qualification' procedures for on-line marking might well occur more rapidly (with markers in centres and without the postal delay) this may compensate for some of the time lost in batching, guillotining and scanning new technology scripts.

One query that was highlighted by the management data concerned the extent to which the new technology system would enhance the utilisation of human resources. While the potentially positive impact of using non-experts to mark non-expert questions was noted, if adjudication demands became too high then there would be the risk that such effects would be attenuated. The fact that around 14% of maths A (response selection) items required adjudication is of some concern in this respect. However, the fact that around 58% of handwriting (expert) items required adjudication raises a potentially larger concern, and this concern would increase as the proportion of double marked expert items increased.

Second, the management data did not provide compelling reason to doubt that accuracy will be enhanced as:

- scripts will be automatically tracked and reconciled through all stages of the external marking system;
- the postal system will be used less frequently (with less chance of loss);
- double marking plus adjudication will improve marking reliability;
- markers will not have to add marks, compute levels, complete mark totals, fill in mark sheets, etc. (eliminating the possibility of careless errors);
- any scripts lost in return to schools could be reproduced from electronic images.

Indeed, the analysis of mistakes arising from conventional clerical tasks (section 5.3.5) made it clear exactly how much error could, potentially, be eliminated from the external marking system simply through the automation of these stages.

Of course, all of these potential speed and accuracy benefits are premised on the assumption that the new technology is designed and implemented to the highest of standards. Were the electronic systems to fail – through lack of functional testing, lack of capacity, lack of technical maintenance or support, incorrect use, lack of back-up, incompatibility, etc. – then any benefits accrued might be eliminated in a flash.

5.4.3 Messages from the measurement data

When considering the measurement data that arose from the Pilot, the crucial question was whether new technology marks were at least as reliable and valid as conventional marks. Unfortunately, as 'correct' marks were not available for any of the items, it was

not possible to address the question of marking quality in a direct manner. However, the analyses did appear to warrant some important conclusions.

First, for the three maths papers, the correlation coefficients between new technology marks and conventional marks were very high. Moreover, as the mean mark differences were low, this suggests that the new technology markers and the conventional markers were awarding very similar marks. This was also true for English spelling markers and for clerical and response selection markers for English reading. This would seem to mutually validate the marks from both systems.

Second, as the correlation coefficients were generally better for response selection and clerical markers than for expert markers (and the mean mark differences were no worse), this suggests that using non-expert markers to mark non-expert questions can be technically effective. Of course, the final non-expert marks included interventions by supervisors and adjudicators. Yet, this kind of intervention was not high for clerical markers (where even the highest adjudication rates were only around 3% of items), suggesting that clerical markers were marking to a high standard. However, intervention was somewhat higher for response selection markers (where the highest adjudication rates were around 14% of items) and further research is needed to explore whether these can be reduced.

Third, as the correlation coefficients for English reading were high for expert, clerical and response selection markers, it appeared that the conventional system and the new technology system resulted in similar rankings of pupil performance. However the mark total for expert reading questions which arose from conventional markers was, on average, two marks higher than that which arose from new technology markers. No clear explanation could be given for this effect and it will be essential to determine whether it can be replicated.

Fourth, across all mean mark differences computed, the mark totals from conventional markers were higher than the mark totals from new technology markers (even if only slightly higher). Once again, no clear explanation could be given for this effect and it will be essential to determine whether it can be replicated.

Fifth, the correlation coefficients for English handwriting and writing were not high. Indeed, the coefficients for writing were lower than would be expected from a marking reliability study that was based exclusively upon conventional markers. Similarly, for handwriting, the typical level of percentage exact agreement between conventional markers and new technology markers was higher than similar figures computed for new

technology markers alone. These results suggested that there may have been specific problems for writing and handwriting, with handwriting in particular appearing to have been less reliably marked by the new technology markers.

Finally, a distinction should be made between potential system effects and potential marker effects. For handwriting, it seems likely that the problems for new technology markers may have been attributable to the relative lack of experience amongst this particular group. This would be to suggest a possible marker effect. On the other hand, the problems could, conceivably, have been due to a more systemic effect; that is, it might simply have been harder to mark handwriting reliably on-screen than on paper.

In contrast, the fact that mean mark differences were always higher for conventional marks – if shown to be a replicable effect – would seem to warrant a systemic explanation that was independent of the specific marker groups employed (particularly bearing in mind that it occurred for all three categories of marker). Exactly what might explain such an effect is another matter though. Importantly, if this effect were shown to be replicable, then it would have implications for scaling up the system. For example, it would argue against employing new technology marking for one region while employing conventional marking for another. Moreover, it might lead to a prediction of a significant drop in the proportion of pupils at any particular level for English if new technology marking were to be introduced nationally. (Note that this drop would be considerably smaller than might be implied by the 20% of pupils mentioned earlier, as these level changes predominantly reflected general marking unreliability, rather than a straightforward decrease in mean marks. That is, level differences in the present study were due not simply to the consistent effect commented upon, but also to the inevitable random errors of marking that are highlighted in any marking reliability study.)

Section 6 An investigation into the frequency of pupil responses located beyond clip image areas

6.1 Introduction

Study 7 investigated the frequency with which pupil responses were located beyond the 'clip image areas' (BtCIA). Such responses would be seen by a conventional marker but not by an e-marker. This could lead to valid responses being missed, or misunderstood, by e-markers, with the consequence of pupils being marked-down inappropriately.

The study assessed the number of occasions in which any part of a pupil's response to an item lay BtCIA. This provided information as to the frequency of responses BtCIA by item and by paper, giving an indication of which items were the most problematic. However, the study did not directly address such other issues as whether the section of the response that was BtCIA would have had an impact upon the mark awarded to the script. This would have required an alternative (and more complex) methodology.

6.2 Method

The investigation revolved around the manual scrutiny of full-page images of pupils' scripts. These were inspected using the NCS Pearson Image Viewer. NCS Pearson were able to superimpose clip image area templates over their full-page images, making manual inspection quite straightforward.

These full-page images for each page of the six test answer booklets were accessed remotely on a dedicated website set up by NCS Pearson using a conventional web browser (Internet Explorer). In practice, there were considerable technical obstacles to be overcome in making these images available to a remote server, both due to security firewall issues at the NFER and problems at NCS Pearson with FTP configuration. NCS Pearson provided considerable support in getting the system linked up and running at NFER, at no additional cost to the project, for which the NFER is very grateful.

Three hundred pupils from 34 schools were sampled, drawn randomly from the population of around 600 pupils that had completed all six tests. In fact, owing to technological obstacles and time constraints, scripts from only 210 of the 300 pupils were assessed to determine whether pupils' responses went outside the clip image areas. For each of the 210 pupils, each of their six test scripts was subject to investigation.

As each script was worked through, page by page, items in which a pupil's response extended outside the clip image perimeter, or in which a meaningful annotation was given

by the pupil outside the clip image area, were recorded on an Excel spreadsheet and data were inputted directly.

Instances of responses BtCIA were recorded on one of six electronic observation schedules (one for each paper). For each paper, each column of the schedule represented a different item and each of 210 rows represented a different pupil. Responses BtCIA were recorded with a cross in the appropriate box. In practice, a certain number of pupils' scripts for each test could not be viewed, due to problems with the underlying image. The number of these inaccessible scripts varied between tests, with the smallest number being 5 scripts inaccessible for mathematics test A, and the largest number being 26 of the mathematics test B scripts. Details of the numbers of unreadable scripts are given in Table 6.1A (Appendix 6.1).

The general logic of the study was to record any instance of a meaningful annotation that strayed even slightly, or was located entirely beyond, the clip image area. This included both responses that were intended explicitly as answers and annotations that may have been recorded for other reasons, such as working in a maths test. Full details of the aspects of pupils' responses which were categorised as 'BtCIA' are given below:

- Any part of a response that was located entirely beyond the clip area.

- Any part of a response that extended even slightly beyond the clip image area and affected the legibility of the response. This included full stops and other punctuation, and the dots of 'i's. It also included portions of letters that extended outside the clip image such that legibility was compromised, for example the descenders of letters such as 'g' or 'y', or the tops of capital letters such as 'T'. However, it did not include instances where the tip of ascenders of letters such as 'h' or 'l' extended beyond the clip image such that the letter in question remained clear. Similarly, instances of letters such as 'e' where adjoiners extended beyond the clip without affecting the clarity of the letter were also not recorded. Although this was sometimes a strict ruling, and the word in question could often have been surmised by a marker viewing the clip, such a stringent definitional framework was necessary to maintain validity and reliability of judgement across more than one researcher, and to avoid excessive complication in a binary data set.

- Any meaningful annotation that was not necessarily part of the intended answer, per se. This included any working for a maths question, including working for the mental arithmetic test. The logic here was that mark schemes explicitly required markers to take such annotations into account.

- Responses to 'please draw' questions for which part of the drawing extended beyond the clip image.
- Any annotation to a graph or drawing.

The following responses were not included as 'BtCIA':

- Any part of a response that extended to the edge of the clip image area, but did not cross it.
- Any marker annotation.
- Any non-meaningful doodle, scribble or irrelevant annotation.
- Responses to 'please circle' questions for which part of the circle extended beyond the clip image, but where it was still obvious which answer was circled.
- Responses to 'please tick' questions for which part of the tick extended beyond the clip image, but where it was still obvious which was ticked.
- A meaningful annotation that had been fully crossed out.

In cases where there was any doubt as to categorisation of a response, responses were recorded as BtCIA.

6.3 Results

Total numbers of responses BtCIA for each question across the sample of 210 pupils are given in Table 6.2A (Appendix 6.1).

6.3.1 Question level analyses

Table 6.1 below summarises the frequency of responses BtCIA for questions from all of the tests evaluated, except the writing test (which had pages rather than questions evaluated for the incidence of responses BtCIA).

Tests with the greatest number of responses BtCIA were the reading test, and the maths A test. The single item with the most responses BtCIA was question 22 of the maths B test, with 48.4% of responses outside the clip area.

189 pupils' writing test scripts were evaluated. 41.3% had a response BtCIA on the first page, 24.9% on the second page, and 2.6% on the third page.

Table 6.1 Frequency of responses BtCIA.

Test	No. of scripts evaluated	Test Questions	
		Questions with 10-20% of responses BtCIA	Questions with above 20% of responses BtCIA
Reading	196	11a (17.3%), 11b (15.3%), 21 (14.8%), 18 (13.3%), 19 (12.2%), 12 (11.2%), 25 (10.7%), 14c (10.2%).	10 (37.8%), 4 (30.6%), 7 (27.6%).
Spelling and Handwriting[1]	195	13 (17.4%), 14 (10.3%).	15 (22.1%).
Maths A	205	22b (19.0%), 24 (13.2%), 1a (13.2%), 15b (12.7%), 15a (11.7%), 7 (11.2%), 9a (10.2%), 22a (10.2%).	16 (42.0%), 11 (33.7%).
Maths B	184	9a (12.5%), 6 (12.0%), , 10a (11.4%), 20 (10.3%).	22 (48.4%).
Arithmetic	197	None[2]	

6.3.2 Pupil level analyses

The same sample of pupils was investigated across all six papers to enable investigation into whether particular pupils were responsible for many of the responses BtCIA, or whether such responses tended to be more randomly spread between pupils and papers. As is apparent from Table 6.3A, 18.1% of the sample presented more than the median number of responses BtCIA across four or more of the tests. This seems to suggest that there are certain pupils (perhaps with poor fine motor control, very large or messy handwriting or a tendency to space their answers badly) for whom there is a tendency to respond BtCIA. A failure to solve the problem of response inaccessibility would affect these pupils particularly.

[1] Of the handwriting scripts, 36.4% had responses BtCIA.

[2] The highest frequency in the arithmetic test was for question 16, with 9.6% of responses outside the clip image.

6.3.3 The relationship with marking reliability

A final analysis considered the question level correlation between frequency of response BtCIA and extent of agreement between conventional and new technology marks. The hypothesis under investigation was whether an inability of on-line markers to see responses BtCIA may have been one cause of their marking disagreements with conventional markers. A high negative correlation coefficient for a paper might support such an inference. That is, if questions with the highest incidence of responses BtCIA also had the lowest agreement between markers, this might be taken as evidence that visibility problems contributed to marking inconsistencies. The correlation coefficients for five of the six papers are presented in Table 6.2 (no coefficient was computed for writing as the paper had only one question).

Table 6.2 Correlation between incidence of BtCIA and marking inconsistency.

	Reading		Spelling		Maths A		Maths B		Arithmetic	
	Kappa	% Agree	Kappa	% Agree	Kappa	% Agree	Kappa	% Agree	Kappa	% Agree
Coeff.	-0.23	-0.17	-0.02	0.01	-0.05	0.08	-0.92	-0.60	-0.06	-0.20
No. qns	31	31	20	20	34	34	34	34	20	20
Prob.	0.210	0.364	0.940	0.974	0.762	0.666	< 0.001	< 0.001	0.799	0.406

It is clear from Table 6.2 that none of the correlation coefficients was significant, with the exception of maths B. Yet, for maths B, the coefficients were actually very high. Although a single 'outlier' had somewhat artificially inflated these figures, even when this was removed the coefficients were -0.45 (Kappa) and -0.61 (Percentage exact agreement).

It is possible, then, that an inability to view responses beyond the clip image areas for the 'calculator allowed' maths paper resulted in marking discrepancies between on-line and conventional markers. On the other hand, it is equally possible that pupils were simply more likely to respond BtCIA for questions that were harder to mark.

6.4 Discussion

6.4.1 General comments

A high proportion (at a rough estimate more than half) of those responses recorded as beyond the clip area would not have endangered the marks awarded as a result of information outside the clip being un-viewable by the e-marker. This would typically be because only a portion of a letter or word was outside, and the response could easily have been inferred. In the case of maths papers – although working was noted BtCIA with considerable frequency – only a few questions explicitly afforded marks for evidence of

correct method. However, it is important to consider the incidence of responses BtCIA from an e-marker's point of view. Even where a response does not extend far outside the clip image area, the marker is typically aware that something has gone over, and has no information as to how far the response might continue. This is clearly frustrating and problematic (see also 3.4.2). Moreover, although working (etc.) might not always be taken into account in practice, there are occasions when it should be in principle, and therefore its visibility is of considerable importance.

It should also be considered that, due to the inherent restrictions in ability range of pupils taking these progress tests, responses were often short, and indeed a higher proportion of questions were not attempted than might have been the case in a sample across the normal range of ability (as in the KS2 tests). From this point of view, the number of responses recorded as outside the clip image area was a good indication of 'danger areas' or item response formats that may need revision in subsequent e-marked tests. However, data from these tests might underestimate the problem.[3]

It is of note that the clip images superimposed onto all the scripts changed position slightly across scripts. This is probably because they were scanned in slightly different positions. As a result, some of the clip images were badly positioned to capture even a well placed response.[4] Certainly the positioning of the clip area is of central importance in obtaining a clear clip, and even a slight change in relative position can have a large effect upon incidence of response BtCIA, as the comments below expound. This issue is therefore of considerable importance as a potential source of inaccuracy in a scaled-up operation.

One rather infrequent but nevertheless important type of response BtCIA was the inclusion of an indication by a pupil that the answer within the box should be read in a different way to that visible within the clip area; for example, an arrow showing that

---

[3] The sample selection for study 7 was drawn from a restricted sub-sample of pupils who had results across all six papers (a total number of just over 600). This could mean that the sample was somewhat biased in its representation of those pupils of the lowest ability, as these pupils were those required to take the Y7PT in both subjects. As such, the sample was not necessarily representative of all pupils who sat the Y7PT tests for the Pilot, and certainly not of the Y7 cohort as a whole.

[4] See for example reading, School 34, pupil 15, p.12, question 20.

responses were to be swapped in their position, or a line altering the order of words in a text or showing that part of a response was located elsewhere on the page.

Specific comments below relate to each test. While reading these comments, it would be advisable to refer to copies of each of the test papers to gain a clearer impression of the implications. Appendix 6.2 presents a script from each of the six papers, marked-up with the clip image area borders and with responses from a pupil.

6.4.2 Reading

Perhaps not surprisingly, those items with responses most commonly extending outside the clip area were those where a long answer was asked for and a relatively small space provided. For both questions 4 and 10, a quotation from the stimulus material was elicited that was perhaps too large for the space provided. A small box was provided as an indication to pupils that only a few words were necessary in the response, but it is reasonable to expect pupils to err on the side of length in such a retrieval task. Thus, for question 4, a common response was 'welcome to the first ever edition of Wildtrack Magazine', when perhaps 'first ever' would have sufficed. Similarly, for question 10, many pupils chose 'I felt rather like an intrepid explorer'.

Perhaps initially more surprising is the high proportion of responses BtCIA for question 7. Why should this question have more than double the number of BtCIA responses than question 18 or 20b, for example? It seems reasonable to propose that this is at least in part due to the fact that question 7 is the first question in the booklet which asks for an extended written response. As such, pupils will be more likely to attempt this question to the best of their ability, or at least to give it a good guess. Also, at the beginning of the test they will have had time enough to express themselves at length and will have had the freshness and confidence to make this expression more likely. As an initial question, it is also easier than the later ones. It is also of note that the perimeter of the clip for question 7 is more tightly close to the right edge of the lines given than in some other questions, such as questions 20a and b. Although the clip perimeter is similarly tight in question 18, it can be argued that answers given here are generally shorter, due to the converse of the reasons given above for the length of an earlier question. Also, question 18 lends itself to short answers such as 'stop killing whales'.

Pupils seemed to have trouble anticipating whether the word they were writing would fit onto the line, and very rarely hyphenated, much preferring to continue a word past the end of the line and beyond the clip area. (However, these responses rarely extended into the grey area of the margin where marks were recorded.) It is worth mentioning that the

layout and format of the tests are not necessarily those which most pupils are used to working with in the classroom. Class jotters generally have lines which extend to the edge of the page, and so it is possible that pupils are seeing available space rather than the end of a stylistically placed line. In a question such as 24, where a boundary to the writing area is clearly demarcated, responses outside this area are far fewer. That answers to 14c, which is similarly boxed, commonly extend outside the clip can be explained by recourse to the most simple reasoning: the answer is too big, and the box is too small. The boundary to the box given in 14a, which is very close to the second writing line, implicitly suggests that it is not a grave sin if a portion of the response should extend slightly outside, as any letter with a descender written on this second line inevitably would do so.

Other noteworthy examples of responses BtCIA were several responses to question 5 where the chosen response text was circled rather than ticked, and one response to question 22 where the connecting lines curved outside the clip area before returning to indicate their connection. Several responses to the multiple choice questions at the beginning of the paper had annotations BtCIA such as 'this one' to show a preference between two choices which had both been indicated.

6.4.3 Spelling and handwriting

The spelling items with higher incidence of responses BtCIA were 13 (17.4%) and 15 (22.1%), which asked for the words 'technique' and 'designed' respectively. Why these two items should have such a high incidence is not immediately clear, but is possibly due to a relatively low positioning of the clip area over the line provided for the response for these items. This may also be linked to the facility of the items themselves, as the hardest words on the script might have been more often crossed out as pupils honed their initial responses to write another attempt above, and thus outside the clip area. This may be supported by the fact that question 13 had the lowest facility across all spelling items (7.57). Question 15 had a less obviously low facility of 28.77. The clip position is therefore more likely to be a cause here.

Again, most of the responses recorded as BtCIA were due to portions of letters extending outside the clip area in such a way as to compromise legibility. This included the dots of 'i's and descenders.

Although the incidence of handwriting scripts with responses BtCIA was quite large at 36.4%, these tended to be due to portions of letters extending beyond the clip area to the right or the bottom of the clip rather than larger portions of words. The handwriting

section of this test is a good example of an item where marks are rarely endangered by a response BtCIA, as not every word needs to be entirely visible for a marker to form an accurate judgement of the overall quality of the response.

6.4.4 Writing

Writing scripts were those which had the highest proportion of 'exceptions' during the scanning process, at around 5%. According to NCS Pearson, this was generally due to scripts for which annotations interfered with the timing and page identity marks in the margins. However, perhaps somewhat anomalously in light of the 5% figure presented above, very few of the writing scripts actually included responses which extended very far outside of the clip image area.

Due to restrictions in the ability range of the sample, few scripts were longer than two pages long, and many only had responses on the first page. This is reflected in the incidence of responses BtCIA on each page (41.3% on page 1, 24.9% on page 2 and 2.6 % on page 3).

The clip image area for the pages of the writing scripts offered pupils a centimetre or so margin to the right of the end of the lines provided, but only around two millimetres below the bottom of the last line. As a result, most of the responses recorded as BtCIA were due to the descenders of letters on the last line of the page. Perhaps one in ten of the writing responses to 'Scene from a Play' had responses BtCIA due to the characters' names being written to the left of the lines provided, with the dialogue then beginning on the line. Perhaps 2 or 3% of the scripts had pictures or other decoration BtCIA, which were not recorded.

It is of note that because the right hand side of the clip area was placed to the right of the end of the lines, many responses which extended past the end of the line were included in the clip image. Indeed, it seems that, on page two particularly, even those responses which did not cross the right boundary of the clip image may have extended to a sufficient extent to interfere with the horizontal timing marks in the right margin of the page. This may explain the apparent anomaly mentioned earlier. Also relevant is the fact that paper markers often placed a tick in the margin of writing scripts or drew a long line at the end of responses; these annotations may well have interfered with the timing marks. Of course, this would not be relevant in a test which was entirely e-marked.

6.4.5 Maths A

Working comprised the majority of responses BtCIA for Test A. Question 11, which required 847 to be divided by 7, and question 16, in which 336 was subtracted from 1025, had the highest incidence of responses outside the clip image area. In both these questions, pupils often performed the necessary calculation outside the clip image area, subsequently writing the answer inside the designated box (and hence within the clip image area).

A large proportion of responses BtCIA for three of the other seven questions with an incidence of responses BtCIA above 10% (1a, 9a, and 24) was also due to working. In the majority of these cases, calculation outside the clip image area would not have affected the pupils' scores for these questions, as marks were not allocated for working per se. However, occasionally a correct answer as part of the working might well have merited a mark.

The responses BtCIA to question 15a were more likely to influence the score given for that question. This question required that pupils respond to each of three statements with either a tick (indicating that the statement was true) or a cross (indicating that the statement was false). Three small boxes (0.8 mm x 0.8 mm) were provided for this response with the clip image area extending approximately 2 mm beyond the left and right hand sides of each box. Pupils who decided to alter their responses almost invariably crossed out their answer within the box and substituted their alternative answer to the right of the box – an area which was not included in the clip image.

Both questions 15b and 22b required the child to explain their response to a previous Yes/No statement in words. Given that only 3 lines were provided for this explanation, and the clip image area was limited to these three lines, it is perhaps not surprising that both of these questions recorded a relatively high incidence of BtCIA responses. These statistics reflect not only inadequate space for children to write their explanations, but also the fact that, in both questions, the clip image extended only slightly beyond the last of the three lines provided for the pupil's response. Consequently, it was almost inevitable that descenders (for example, the letters 'y' or 'g') used on this line would not be completely included in the clip image area.

6.4.6 Maths B

Again, the great majority of responses BtCIA on this paper were due to pupils' rough working outside the answer box provided. The item with by far the largest incidence of

response BtCIA was 22, with 48.4%. This was principally due to pupils completing the sequence provided on the dots given with the number 19, although this was not part of the marked response. This question was also the only item in the paper which asked for a written explanation, and the clip area was quite tightly close to the right end of the lines provided.

Items 6, 9a, 10a and 20 also had an incidence of response BtCIA higher than 10%. In most cases these were due to working outside of the answer box. Although question 10a provided a large box for working, many pupils had not grasped the principle of multiplication necessary to answer this question, and attempted to solve it through lengthy addition. Many of the responses to question 20 which were BtCIA were due to annotations or sketches on or beside the diagram provided.

Many responses to question 17 included an arrow with a 'tail' which extended outside the clip perimeter. These were not recorded (in accordance with the agreed specification).

It is of note that some responses to question 13 used lines to indicate the position of the numbers provided in the Venn diagram rather than writing them again in the applicable place. This is a good example of a response that could merit marks and would be clear to a paper marker, but entirely outside the image given to an e-marker.

6.4.7 Arithmetic

The relatively low incidence of BtCIA responses in the mental arithmetic test may be attributed to several causes. Firstly, the verbal instructions accompanying the test specifically requested that pupils make their responses to each question inside the appropriate box ("On your sheet there is an answer box for each question, where you should write the answer to that question and nothing else"). Secondly, unlike, for example, question 15b in maths test A, there was sufficient space within each answer box for pupils to change their answers without the new answer extending beyond the clip image area. Thirdly, time constraints of the test afforded little time for working. This particular test allowed a 5, 10, or 15 second response time for each question, reducing the likelihood that pupils would attempt to calculate their answers in the area outside the clip image. Indeed, verbal instructions prior to the test stated "Do not try to write down your calculations because this will waste time". Pupils were also told that they should work out their answers to each question in their heads, although they could "jot things down" outside the answer box if necessary. Almost all of the responses BtCIA in this test, including most of those for question 16, which recorded the highest percentage of responses BtCIA, reflected this type of working. The extension of part of an answer (for

example, the lowest quarter of the number 8) slightly beyond the clip image area was the next most common response BtCIA. However, this type of response occurred relatively infrequently.

It is likely that the limited time available to calculate/alter responses is a major cause of the small number of responses BtCIA in the mental arithmetic test. Nevertheless, adoption of the explicit verbal instructions used in this test, which specify that answers should be written within the designated boxes, and perhaps also a warning that answers written outside this area will not be marked, as well as the provision of adequately-sized boxes, might help reduce the incidence of responses BtCIA in the other Y7PTs.

6.5 Summary and conclusions

The frequency of pupil responses which extend outside the clip image area was considerable. With the exception of the mental arithmetic test, each test evaluated had at least one item for which the incidence of response BtCIA was greater than a third of those pupils taking the test. Although it is to be considered that many of these responses were not seriously compromised in their legibility, or did not afford information that might have merited a change in the e-markers' evaluation of the mark to be awarded, there can be no doubt that at such a high rate of incidence there were a significant number of responses in which these issues would have been relevant.

As such, the problem of responses or annotations outside the clip image area presents a threat to e-marking validity in a scaled-up operation.

Common reasons for pupils' responses extending BtCIA varied between tests. Across all tests, some pupils provided answers which were subsequently crossed out. The new answers were often written outside the clip areas. In the reading test, it was common for written responses to continue beyond the end of the line provided, and even above or below the designated area. These issues also applied to the spelling and handwriting test and to the writing test scripts. It is also of note that responses to the writing test often continued past the end of the line into an area where the timing or identity marks of the page might have been interfered with.

In the maths tests, responses BtCIA were usually due to rough working. Often, the clip area provided was rather small, showing only the answer box provided rather than also including the accompanying space.

It seems clear that one expedient solution to the problem of responses outside the clip areas would be simply to have clip images considerably larger than the spaces provided in

the booklets for the answers. A clip which extended a few centimetres further in both dimensions would have encompassed the great majority of those responses recorded, certainly in the case of the English tests. Similarly, to afford markers access to the full page images in instances of uncertainty would ensure that no aspect of a response was unavailable to a marker. The technical implications of such changes – presumably in relation to cost and speed – need to be explored with future contractors.

In addition, to help prevent such a high incidence of response BtCIA, there might be scope for warning pupils in advance of the importance of writing in or very near the space provided, and of not tampering with timing or identity marks. Indeed, the clip image areas might even be subtly transposed onto the scripts. Finally, in cases where a written response is asked for (specifically some of the reading questions and the writing scripts) margins might be more clearly indicated with a thick line, and more space provided.

Section 7 An evaluation of the 2001 New Technologies Pilot

The preceding sections have presented results and insights from the seven studies of the Evaluation. The purpose of this last section is to present a synthesis that will achieve the four broad objectives listed in the Introduction. Encapsulating the more detailed specifications of the Project Initiation Document, these four objectives were:

1. to evaluate whether the NTP contractor managed successfully to implement the agreed procedures for 2001;

2. to evaluate whether the procedures implemented during 2001 were effective in delivering significant benefits without undue costs;

3. to consider whether the procedures implemented during 2001 might be scaled-up for all national curriculum tests to deliver significant benefits without undue costs;

4. to consider whether revised procedures for future years might deliver significant benefits without undue costs.

Section 7 begins by considering the success of NCS Pearson in implementing new technology marking for the Year 7 progress test scripts sampled during the Pilot. The Invest to Save Budget submission anticipated that on-line marking might initially be offered at dedicated marking centres and, over time, move toward home delivery with training and supervision conducted remotely via the web. The Pilot adopted the former approach, being based in a single centre, with training and supervision being conducted largely face-to-face. As such, the following discussion focuses primarily upon the extent to which the centre-based on-line marking model was effective on a small-scale, and upon its potential for scaling up to a national level. Attention is then turned to the potential for realising a web-based on-line marking model. Section 7 ends with conclusions and recommendations for future action.

7.1 The success of the Pilot contractor

The remit of the Evaluation was to establish the success of the Pilot and the first question that arose was whether the Pilot contractor had managed to deliver the new technology as promised. As explained in Section 2, NCS Pearson did achieve this goal with a considerable degree of success. The key features of their system were as follows:

- advance script personalisation and electronic tracking through all processing stages;

- script batching, guillotining and scanning;
- electronic storage of both full-page script images and item-level 'clip' images;
- categorisation of questions as requiring unskilled, semi-skilled or expert marking;
- electronic distribution of item clips to unskilled, semi-skilled and expert markers;
- centre-based on-line marking;
- face-to-face supervision (with an on-line component);
- double marking for unskilled and semi-skilled markers;
- single marking, plus monitoring, for expert markers;
- adjudication, by senior markers, of discrepancies arising from double marked items;
- production and analysis of management and measurement data;
- automatic aggregation of marks and allocation of levels;
- production of detailed reports on pupil performance for schools, prepared electronically and accompanied by electronic script images.

7.1.1 Shortcomings of the Pilot

In delivering the new technology system, there were three factors that compromised the extent to which general conclusions could be drawn from the Pilot:

- the failure to recruit a sufficient number of expert markers with prior experience of conventional marking;
- the failure to scan and mark the full complement of scripts;
- the use of Netgrade rather than the full ePEN software.

Although the first two were unfortunate, they appeared to have been largely beyond the control of NCS Pearson. NCS Pearson explained that Netgrade performs the same function as the full ePEN software (which they propose to use in subsequent projects); however, the fact that the full system was not implemented meant that there was no evidence to confirm the nature or quality of ePEN.

There were also functions and procedures within the NCS Pearson system that either failed, or that appeared sub-optimal for the purpose of on-line marking. While these did not generally compromise the Pilot, they constituted specific weaknesses that would need addressing in future pilots or trials. They included:

- a break-down in the on-line supervision processes;
- certain frustrating software peculiarities;
- a failure of the absolute-marker reliability ('validity') function;
- an apparent lack of attention to the possibility of incorrect mark entry (through mis-selection of mark options).

Procedural concerns were also raised with the question allocation process, the (apparent lack of) procedure for verifying pupil names on scripts, and the implementation of the on-line mark scheme. Additionally, around a third of expert markers would have appreciated extra training and the supervising markers did not always feel well prepared for their tasks.

Despite these reservations, it should be reiterated that the Pilot was generally very successful and demonstrated that on-line marking is a real possibility for national curriculum tests.

7.1.2 Lessons learned from the Pilot

Finally, there were two specific areas of concern for the Evaluation team that highlighted aspects of the UK system to which the new technology may have to adapt. These should not necessarily be considered failures on behalf of the contractor; instead, they are lessons to be learned for future projects:

- the frequency with which pupils' responses were located beyond the clip image areas;
- the specific supervisory needs associated with single (rather than double) marking.

7.1.2.1 Responses located beyond the clip image areas

Although NCS Pearson stored both full-page and item-level clip images, only the latter were distributed to markers for marking. Moreover, as each clip image area was no larger than the space in which pupils were intended to respond, any responses located beyond those spaces were not available to markers for scrutiny. Of course, where a marker was

unable to see a response, s/he was unable to award a mark for it. Although many of the responses located beyond the clip image areas were no more than the extensions of letters or words (which could be inferred) other responses were not. Sometimes entire answers lay beyond the clip image areas (particularly when responses had been crossed out). On other occasions mathematical working or annotations to graphs or pictures were out of sight. While these would not necessarily have determined the mark awarded, they might still have influenced it. Indeed, for the maths B paper, there was a significant correlation between the prevalence of responses located beyond the clip image areas and the extent of disagreement between conventional and new technology markers. That is, the inability to scrutinise responses located beyond the clip image areas may have had a significant impact upon marks awarded (although other explanations are certainly possible).

This problem will need to be addressed somehow. Markers cannot be expected simply to 'flag' items where they suspect responses are out of image as, in many cases, they will simply not know. Nor can it be assumed that double marking will overcome this problem as, if a response is invisible to one marker, it will be invisible to both. Markers in the Pilot were not happy being required to mark only what they saw (when they suspected that pupils had written more) because they felt that this was not fair – and this was despite realising that pupils would not receive feedback from the Pilot! If pupils cannot be required to write only in the intended spaces then perhaps markers should be able to access full-page images? Or, perhaps they might be provided with half-page images, with light shading beyond the intended clip image areas (to de-emphasise, but not to obscure, other questions and responses)?

7.1.2.2 Models of supervision

The present UK external marking system evolved to support single (rather than double) marking. The new technology marking system, on the other hand, was optimised for 100% double marking. Within the Pilot, the basic supervisory model for double marking (of clerical and response selection markers) was fairly straightforward.[1] Markers were trained to a sufficiently high standard and then they began marking. When they made mistakes these were detected (as mark-re-mark discrepancies) and corrected. If markers had made too many mistakes they could always have been stopped. However, the failure of an individual marker was ultimately not a threat to the final product. Thus, the

[1] This sketch only highlights the major components of the supervisory model (ignoring relatively minor issues such as 'sticky notes') and makes no formal distinction between supervision and adjudication.

emphasis of supervision for on-line double marking was upon quality control (rather than quality assurance).

With single marking the situation is different because the failure of an individual marker will ultimately compromise the final product. Indeed, every single item that s/he marks incorrectly will compromise the final product to some extent. During conventional marking this is addressed through sampling stages in which markers must mark a set number of scripts within specified absolute mark difference tolerances:

a. first sample (a post-training and post-standardisation 'certification' stage, using live scripts, that must be completed satisfactorily before marking proper commences);

b. second sample (a repeat of the above stage for those who fail to mark within tolerance the first time around);

c. final sample (undertaken around two-thirds of the way into the marking process to ensure that the correct standard is still being applied).

This approach is far from ideal and does not eliminate all subsequent error. However, it does illustrate that for single marking the principal supervisory function is to ensure competence at the outset, while the secondary supervisory function is to ensure that it is subsequently being maintained. Thus, the emphasis of supervision for conventional single marking is upon quality assurance (rather than quality control): helping markers to identify and overcome their mistakes and misunderstandings early on, thus preventing (rather than managing) error.

Now, although NCS Pearson provided a wealth of monitoring functions for the supervision of expert markers (the only category of marker to single mark), these functions were not designed to support quality assurance. Indeed, being based upon random sampling principles, they were more appropriately designed to support quality control for accountability or audit. That is, by the time a marker had marked a sufficient corpus of monitoring items to enable valid inferences to be drawn concerning her marking quality, any damage would already have been done for a large number of pupils. Had the monitoring instead been concentrated at the commencement of marking for each question then any problems could have been identified immediately and the necessary support given.

This discussion is not intended as a criticism of NCS Pearson, as QCA appeared not to have specified any alternative supervisory models. Instead, it is intended to highlight

some very major decisions that will need to be made concerning the management, training, 'certification', supervision and monitoring of the marking hierarchy. All of these decisions will ultimately be affected by the decision whether to employ single or double marking for expert questions.

If it were decided to employ a 100% double marking model for expert markers then there would be little reason to depart significantly from the approach successfully adopted for clerical and response selection markers during the Pilot. However, this would increase the number of expert markers required by the system. Although expert markers would already have been freed by the employment of response selection and clerical markers, results from the Pilot suggested that the expert marking load is still around three times that of other marking categories (so the number of experts freed might not be as many as anticipated). Furthermore, double marking would not simply mean many more expert markers, it would also mean many more supervisors for adjudication. Particularly for English writing and handwriting, mark-re-mark discrepancies are the norm, rather than the exception. One might question the feasibility (let alone the logic) of a senior marker adjudicating two expert markers for around one in every two items marked.[2] Nevertheless, the payoff from implementing a 100% double marking model across all marking categories would be very considerable indeed, at least in terms of marking reliability. There is much to commend this model.

If, on the other hand, it were decided to employ a single marking model for expert markers then it would be necessary to develop a new supervisory model to support it. This will not necessarily involve random sampling (to a fixed percentage) throughout the duration of the marking period. Indeed, as suggested above, there are good reasons for not adopting such a model for single marking. Certainly there would have to be more emphasis upon initial training and competency 'certification' (akin to the conventional sampling procedures). The extent to which this would focus upon all expert questions at one time, or upon individual questions as and when appropriate, would be a matter for consideration. Likewise, the extent to which this could be automated should also be considered. Most importantly, though, it will be crucial to determine the role of supervisors in giving formative feedback to markers. This will have capacity

[2] Of course, there are alternatives. For example, only items with discrepancies at or above a certain AMD might be sent for adjudication, while items with discrepancies below that AMD might receive an average of the two expert marks.

implications, as the need for such input is likely to be particularly high at the commencement of the marking period. To the extent that feedback for on-line markers at a marking centre would need to be very rapid (in contrast to feedback on the conventional first sample) this might prove problematic.

7.2 Potential benefits, risks and costs associated with the centre-based on-line marking model

The web-based on-line marking model is very ambitious and it may not be possible to deliver it all in one go. Indeed, it may simply not be feasible in the short- to medium-term. During the Pilot, only a centre-based model was implemented and the following discussion will consider the extent to which it was successful and the potential for scaling up to a national level. The centre-based model brought markers together, within a dedicated marking centre, to mark across an intranet. Instead of being supported primarily on-line, most of the support that they received from administrators and senior markers was face-to-face.

Generally speaking, the Pilot was successful and did not obviously deny the feasibility of a national centre-based marking model. The following sub-sections will discuss the advantages of this model over the conventional home-based paper marking model.

7.2.1 The potential benefits of a national centre-based on-line marking model

7.2.1.1 Speed

There are major procedural differences between conventional and new technology marking and there are strong *a priori* reasons for thinking that these differences may result in a reduction in the duration of the external marking process. As explained in Section 5, the principal reasons are as follows:

- teachers will not have to sort completed scripts into order (although they may spend more time distributing personalised papers to pupils prior to each test);
- the postal system will be used less frequently;
- the pool of markers can be increased as non-experts are drafted in to mark non-expert questions (although the higher the proportion of double marking for experts, the smaller the potential time gain would be);

- markers will not have to add marks, compute levels, complete mark totals, fill in mark sheets, etc. (and checks on these procedures will be eliminated);
- data will be input directly by markers and not subsequently by inputters;
- assuming continued full-page image scanning, any review procedure could be initiated by telephone as the script would not need to be sent by post.

There was no compelling evidence from the Pilot to conclude that these procedural differences will not lead to time savings in a national centre-based on-line marking model. On the other hand, there was no compelling evidence from the Pilot to conclude that markers would actually mark more rapidly using the new technology apparatus. Thus, the savings in elapsed time would appear to come predominantly from the elimination of significant conventional stages. A larger pool of markers might also contribute to time savings, although the positive impact would be reduced if greater emphasis were to be placed upon double marking.

Of course, the on-line marking model does not simply eliminate unnecessary conventional stages, it also adds new ones. In particular, there are batching, guillotining and scanning stages that have to be undertaken before any marking can begin at all. Clearly, then, time that is gained by the elimination of certain conventional stages might be lost through the introduction of additional new technology stages. The Pilot was clearly delayed by the inability to scan a sufficient number of scripts to keep the early stages running smoothly. However, there are grounds for attributing this problem predominantly to conventional markers not returning scripts on time. Further piloting, in optimal circumstances, will be necessary before it will be possible to determine whether or not the batching, guillotining and scanning stages pose a significant threat to speed of the on-line marking system. In particular, it will be important to monitor rates of scanning 'exceptions' and 'attachments' (and the potential knock-on delays from small numbers of scripts for large numbers of batches).

7.2.1.2 Accuracy

Just as certain procedural changes may lead to savings in elapsed time, it is likely that others may enhance the accuracy of the external marking system – both in terms of processing (e.g., more accurate script tracking) and in terms of product (e.g., more accurate marks). These changes include:

- scripts will be automatically tracked and reconciled through all stages of the external marking system;
- the postal system will be used less frequently (with less chance of loss);
- double marking plus adjudication will improve marking reliability;
- markers will not have to add marks, compute levels, complete mark totals, fill in mark sheets, etc. (eliminating the possibility of clerical errors);
- any scripts lost in return to schools could be reproduced from electronic images.

Once again, there was no compelling evidence from the Pilot to conclude that these procedural differences will not lead to improvements in accuracy. Of course, there are other novel aspects of on-line marking that might threaten the accuracy of the system, and these predominantly revolve around the new technology. Were these systems to fail – through lack of functional testing, lack of capacity, lack of technical maintenance or support, incorrect use, lack of back-up, incompatibility, etc. – then any benefits accrued might be eliminated in a flash.

The analysis of measurement data from the Pilot was extensive and the comparison between marks from conventional and new technology procedures was very informative. While there was no definitive evidence that one system produced more reliable and valid results than the other, there was evidence that mutually validated both of the systems, particularly for the maths papers, but also for English reading and spelling. English writing and handwriting resulted in lower levels of agreement than might have been hoped for, yet this lack of mutual validation might simply have been due to the relative lack of marking experience amongst the on-line markers.

One of the most important conclusions from the Pilot was that marks arising from unskilled (response selection) and semi-skilled (clerical) on-line markers – once adjudicated – were very similar to those arising from expert conventional markers. This is further boosted by the finding that clerical markers tended not to require a great deal of adjudication anyhow. These results support the use of non-experts for marking less complex questions.

The most peculiar result to emerge from the measurement data was that all of the mean mark differences computed between conventional and new technology marks were positive (meaning that conventional marks were, on average, higher than new technology marks). For English reading an exceptional difference of two marks was observed. When

translated into levels for English overall, it meant that 20% of pupils would have received a lower level had the new technology marks been used instead of the conventional ones (while only 7% would have received a higher level). There was no clear explanation for this finding and further research will be crucial to determine whether it represents a genuine systemic difference. If this effect were shown to be replicable, then it would have implications for scaling up the system. For example, it would argue against employing new technology marking for one region while employing conventional marking for another. Moreover, it might lead to a prediction of a significant drop in the proportion of pupils at any particular level for English if new technology marking were to be introduced nationally. (Note that this drop would be considerably smaller than might be implied by the 20% of pupils mentioned above, as these level changes predominantly reflected the consequences of marking unreliability, rather than a straightforward decrease in mean marks.)

New technology offers the potential to enhance the accuracy of marks awarded during external marking. In particular, the simplicity and speed with which responses can be double marked is a major opportunity. There are also other significant benefits, though, even if single marking were to be retained. For example, the distributed marking system means that the impact of a 'rogue' marker would be spread across many pupils and schools; this would be likely to eliminate any apparent need for school (R3) reviews and to reduce the number of pupil (R2) reviews – even if the quality of external marking did not improve overall. In addition, the distributed marking system means that markers cannot mark responses to individual questions in the context of responses to previous questions (i.e., they will not be tempted to give candidates the benefit of the doubt on the basis of previous correct responses).

As a final word of caution, though: in the desire to embrace the potential of on-line marking, it will be important to ensure that particular strengths of conventional marking are not over-looked. For example, during the Pilot there was some concern that markers were unable to highlight items that they were unsure about and wished to return to later. This is a normal and crucial component of conventional marking and supports a quality assurance, as well as a quality control, function: markers discuss these problematic items with supervisors and are able to modify their marking in the light of feedback. Yet it is easy to see why this might be somewhat incompatible with the on-line model: as long as an item is highlighted for later consideration it is no longer being processed through the system and this will have knock-on effects for functions like adjudication and progress

monitoring. Exactly how to incorporate strengths of the conventional system into the new technology system will be a challenge for future pilots and trials.

7.2.1.3 Management and administrative systems

The Invest to Save Budget submission highlighted speed and accuracy as the most immediate potential benefits from the implementation of new technology. Yet it also proposed improvements in management and administrative systems, citing increased flexibility, improved operability, a more automated service, and a more seamless approach across all parties involved.

Indeed, there would seem to be two major management and administrative advantages of a national centre-based on-line marking model. First, it would combine two functions that are presently operated separately: external marking and national data collection. Second, as all major procedures would be managed via central computer systems, the capacity to monitor and review all processing stages would be greatly enhanced.

Bringing marking and data collection together within a single system, managed by a single provider, would seem to be a major advance. It is hard to see how this could not improve operability and result in a more seamless approach. The various stakeholders interviewed for Section 4 stressed the problems that can occur when multiple contractors are required to collaborate using procedures and systems that are not necessarily entirely compatible.

Likewise, the potential for providing more management information upon all stages of the external marking process should improve operability; and the general move towards electronic processing will maximise the potential for interfacing with the DfES Information Management Strategy. On the other hand, there are significant risks associated with the potential for producing vast quantities of management data. Perhaps the most significant risk is that data production may initially be seen as a peripheral concern, that it may be left to evolve erratically and that the resultant information systems would be ineffective (due to incompleteness, fragmentation, incompatibility, inaccessibility, etc.). New technology offers the potential to provide a wealth of data supporting information functions that have not previously been possible. To realise these functions optimally will require detailed planning to determine exactly what kind of information is required at which stages for what purposes; it will require expert technical advice to facilitate the generation of data that will accurately support these information

functions; and it will require the training of staff to understand and use the information provided.

7.2.1.4 Stakeholder satisfaction

The Invest to Save Budget submission also specified the maintenance of confidence in the testing process as a potential advantage of the adoption of new technology. This can be interpreted more widely as improving levels of satisfaction amongst all stakeholders of the external marking system. This would include parents and pupils, teachers, unions, markers, service providers, service contractors and governmental agencies.

7.2.1.4.1 The general public, parents and pupils

If it were successful in improving the accuracy of external marking then a national centre-based on-line marking model would be likely to improve levels of satisfaction amongst the general public, parents and pupils. However, it would be important for QCA to ensure that the public were not convinced by sceptics that any improvements in marking accuracy were being achieved at the expense of validity, due to scripts being 'marked by computers'. This may prove to be a sensitive public relations issue. It may need to be ensured that the new marking model is transparent, open and widely understood.

There might also be a threat to public confidence if the on-line system were seen to down-grade the professionalism of external marking. Some believe that public trust in the conventional model is largely due to the central role of professional teachers. If certain marking functions were seen to be taken over by unskilled, or semi-skilled staff, this might have adverse consequences for public confidence.

7.2.1.4.2 Teachers and unions

Levels of satisfaction amongst teachers would be likely to improve if accuracy and speed improved, particularly if teachers saw levels of review requests decrease and if results from national curriculum tests could be finalised before the end of the summer term. In addition, any decrease in administrative burden would be appreciated. Not having to sort scripts into alphabetical order is one such advantage and there are potential advantages with respect to pupil registration. Finally, to the extent that more detailed information on pupil performance will be supported by the new technology system, teachers' levels of satisfaction are likely to increase. However, it will be important to ensure that teachers receive useful additional information rather than an impenetrable data overload. Moreover, it will be important to determine whether teachers feel that the loss of personal

feedback from a specific marker is a price worth paying for the new kinds of performance information that they stand to gain.

Whether the professional associations and unions might raise concerns is somewhat hard to predict. One potentially contentious issue is that of contracting-out a high-profile and politically sensitive component of national curriculum assessment to an exclusively 'for-profits' company. The extent to which this might be perceived as a threat to the integrity of the system is not clear.

7.2.1.4.3 Markers

The most important hearts to win are those of the markers. If markers cannot be convinced to co-operate then a national centre-based on-line marking model could be a disaster of epic proportion. Some insight into markers' impressions was gained from the Evaluation. However, the research was on a very small scale and the markers were not representative. Ongoing monitoring of marker attitude and behaviour will be essential to the success of the new technology programme.

The new technology principle

Generally speaking, the feeling from expert markers in the Pilot was that on-line marking was the way forward; they welcomed not having to deal with piles of paper, administration, mark totalling and packaging of scripts. Indeed, they were slightly more positive after having experienced the new technology system than before. Many of the markers expressed particular satisfaction with being able to build up an expertise on a single question and found this approach quick, enjoyable and interesting.

On-line experiences

While markers found the Pilot an enjoyable experience, a number of specific concerns were registered. Delays were not appreciated, especially as they had given up valuable time to undertake the work. Also, although the software was not complex to operate, the terminology of many of the functions was considered to be poor. As mentioned previously, markers did not like being unable to see beyond the clip image areas when they knew that responses had been written there. Further, one of the principal unfavourable comparisons with conventional marking was not being able to return to awkward items (other than those most recently marked). Markers felt that this made their marking look less reliable than it really was. Finally, some complained of physical aches

and pains in the back, eye, head and wrist. This is a Health and Safety issue that it will be important for QCA and any contractor to consider thoroughly.

Centre-based marking

Markers appreciated the social element of centre-based marking (although whether it would have been quite so sociable if the workload had been as intended is another matter). The only negative comment about the marking environment was concern from a few markers over the distracting noises of scanning machines and chatter.

The most important message, though, was that the majority of markers would have preferred to have worked at home. Those who had long distances to travel to the centre (particularly those who travelled each day) did not enjoy their journeys. Working at home was equated with working at their own pace, in peace and quiet; moreover, it meant not having to leave children during holidays. There was a general feeling that, if a centre-based model was implemented, it would be better if markers did not have to travel far.

It is worth noting results from the 2000 KS2/3 Marking Evaluation Report. Most of the 2000 markers were full-time or part-time teachers who marked predominantly during the evening or at weekends. Requiring markers to travel to marking centres to mark during weekdays would constitute a significant culture change. This would seem to be a major risk – perhaps *the* major risk – for scaling up to a national centre-based on-line marking model.

7.2.1.4.4 Suppliers, contractors and agencies

Clearly, the suppliers, contractors and various agencies would be satisfied if the various potential benefits discussed throughout 7.2.1 could all be realised. All of the representatives interviewed for Section 4 felt that there was a substantial need for the kind of improvements in external marking that the new technology promises.

7.2.1.5 Financial viability

The final potential benefit of a national centre-based on-line marking model is that it will reduce costs associated with external marking, which are presently very high. It was not within the remit of the Evaluation team to provide a detailed analysis of whether new systems will provide good value for money. However, it is worth noting a few of the major issues for consideration.

The costings of the Invest to Save Budget submission were largely premised on a web-based scaling up in which savings are due to:

1. reduced registration costs (school burden, administration)

2. reduced training costs (administration, organisation, attendance, travel, preparation)

3. reduced collection costs (school burden, marker burden, postage);

4. reduced marking costs (marking, clerical checking);

5. reduced management costs (through greater competition).

Reducing registration costs assumes that schools become equipped with good quality on-line computer technology, well maintained pupil records and staff competent to run automated registration procedures. Only time will tell.

Reducing training costs assumes that markers will be trained remotely via the web and will not have to travel to training centres. However, markers will still need to be paid for any training undertaken. More importantly, there will be no training cost reduction if on-line marking takes place in centres rather than at home, as similar expenses will be incurred.

Reducing collection costs assumes that the annual fee for renting software and hardware (and associated technical, managerial and support staff) for batching, guillotining, scanning, marking, processing, etc. is more than offset by savings in school burden, marker burden, postage and data entry. This would seem to be the crucial element of the financial case. Unfortunately, though, it is far from clear whether the net impact would be positive or negative.

Reducing marking costs assumes the elimination of clerical tasks (around 10% of the marking fee). However, as the marking load will be doubled for clerical and response selection questions (and there is likely to be double marking for at least a proportion of expert questions), there will be considerably more marking taking place and this seems likely to increase marking costs.

Finally, reducing management costs assumes that savings will accrue through opening the market up to competition and by contracting more suppliers to undertake less complex work packages. However, it is not at all clear that more suppliers would be required, or

desirable. Nor is it clear that there will be a sufficient number of competent potential contractors for the market-forces model to apply effectively.

In conclusion, the argument that implementing new technology will result in the projected financial savings is not entirely persuasive. Indeed, there would seem to be a significant risk of increasing the costs of external marking. Amongst the stakeholders interviewed for Section 4, only NCS Pearson representatives were confident that the new technology programme would result in significant financial savings.

Scaling up to a national centre-based on-line marking model would introduce specific additional costs, relating to marker travel, accommodation and administration. These costs might even exceed marking fees if only a few large centres were used (and considerable travel was required). Further costs would be incurred in renting and staffing the marking centres and in setting up the necessary hardware and software systems.

7.2.2 Additional issues to be addressed in scaling up to a national centre-based on-line marking model

7.2.2.1 Marking centre models

The first step in establishing the viability of a national centre-based on-line marking model, would to be determine the most appropriate marking centre structure. Three possible models are suggested below:

1. large, regional, custom-built centres, owned by the QCA (e.g., QCA marking centres);
2. small, local, custom-built centres, rented by the QCA (e.g., DVLA driving test centres);
3. small, local, adapted centres, rented by the QCA (e.g., LEA or school IT centres).

Large custom-built centres would be ideal if the set-up costs were not prohibitive, if markers were prepared to travel to them, and if travel and accommodation costs would not also be prohibitive. Unfortunately, all of these seem unlikely and, as the centres would be used for only a short period each year, this option does not seem very promising.

Small custom-built centres, such as those already used for driving tests, are a possibility. However, it is likely that they would only be available during certain periods (i.e., over night or on Sundays). If markers would not accept these periods then the model would fail.

The same might also be true for small IT centres that would have to be adapted, such as LEA or school IT centres. However, holidays and understanding suppliers might make rental during acceptable working periods more realistic. On the other hand, this approach would introduce considerable task inconsistency into the marking model, as markers would be working in substantially different environments, with substantially different hardware. The software roll-out would also pose significant obstacles. In addition, there would be security issues if lots of small centres were used.

Finally, the issue of where to locate the script scanning centre(s) requires further deliberation. In principle, scanning could occur in a single location. However, a regional network of centres would help to spread load and minimise risk.

7.2.2.2 Marker appointment, training, supervision and allocation

The appointment of markers presented a problem during the Pilot. It would be important to determine that enough markers could be appointed for scaling up to a national level. Particularly if the marking were concentrated in a small number of marking centres, would it be straightforward to appoint a sufficient number of temporary, unskilled or semi-skilled workers during the marking periods? Would expert markers be prepared to travel (particularly if only to mark for a couple of hours in the evening)? What additional costs would be incurred through the appointment processes (particularly if a larger number of regional centres were used)?

As discussed earlier, there are major issues of training, standardisation, 'certification', supervision and monitoring to be decided. These decisions will influence subsequent decisions concerning how to staff the marking centres. It is presumed that marking hierarchies would operate in each marking centre. However, no assumptions concerning the nature of these hierarchies are made, as these may have to change to accommodate new supervisory practices.

The issue of work allocation will also have to be given further thought. This is closely linked to another thorny issue of how to manage marker pay. If markers were paid by the hour, as would seem appropriate in a centre-based model, it would need to be determined whether this would require fixed period contracts. If so, then contingency plans would be necessary to accommodate possible work delays that reduced the number of items marked within the set period (as occurred during the Pilot). New payment models will need to be piloted and trialled alongside other aspects of the new technology system.

7.2.2.3 Special arrangements

If scaled to a national level, it would be essential to ensure that none of the various 'exceptional' circumstances was overlooked. The system must be sufficiently flexible to deal with arrangements for amanuenses, enlarged test papers, modified large print, Braille, transcripts, etc..

7.2.3 Additional risks and potential costs associated with a national centre-based on-line marking model

Additional threats to a national centre-based on-line marking model are presented below under three headings: management threats; technology threats; and assessment threats.

7.2.3.1 Management threats

A general threat identified within the Invest to Save Budget submission was that management would not keep on top of technological advances and the eventual systems would soon be obsolete. In a rapidly changing technological environment it is good to keep this caution to the fore. Not to foresee this kind of a programme failure would be very embarrassing for all involved. Similarly embarrassing would be failure due to incompatibility between the new technology system and other developing projects, programmes or strategies, either from within the QCA or from without. Links to projects of obvious relevance need to be formally established and maintained so as to keep abreast of changing specifications.

Then there are more specific management threats associated with the role of the QCA in the development programme. Risks of implementation failure increase significantly as time constraints become more pressing, yet QCA is not renowned for prompt decision making. Decisions need to be made as early as possible and should leave sufficient time for both completion and testing. Implementation also becomes more complex as more contractors are involved. There was a feeling from certain of the stakeholders of Section 4 that the QCA needs to improve its approach to planning and contracting, as cross-contractor co-ordination has not always been optimal in the past.

As mentioned in previous sub-sections, the new technology roll-out will necessitate the appointment of staff to marking, administration, support and management positions. Neither NCS Pearson, QCA nor any of the awarding bodies have sufficient human resources to take over these functions at present. If the system were to scale to a national

level, there would be significant training and development demands to support new layers of personnel.

Finally, there are potential threats that arise from explicitly encouraging private sector involvement in areas that have traditionally been dominated by non-profit-making educational organisations. First, there is the risk that organisations with established educational credentials and effective structures for managing external marking personnel and procedures (i.e., awarding bodies) may be too quickly sidelined in favour of companies with limited educational knowledge and experience but with the requisite technological capacity. Second, there is a risk that private companies – driven ultimately by profit-margin concerns – might raise threats to the wider national interest; for example, if unanticipated copyright controversies were to compromise the implementation of new technology systems in years to come.

7.2.3.2 Technology threats

All participants in the interviews of Section 4 spontaneously mentioned the 1998 score sheets "disaster". In recent years, both QCA and SQA have had to deal with the impact of large-scale technological failure. Indeed, when technology fails, it has a tendency of failing on a large scale and this can have very serious consequences for organisations as they are dragged through the media mangle.

Technological failure can have such wide-spread impacts when automation results in the replacement of many independently functioning units with a single one. Thus, instead of hundreds of markers computing mark totals for their own scripts, a single computer takes their place using a single algorithm. Similarly, every single stage of the new technology processing model would rely on a single software system; script scanning would rely on a small number of intricate and expensive machines; complex item generation and distribution routines, mark aggregation algorithms and codes for producing valid statistics would rely on small teams of programmers; and the entire model assumes that the net supporting the distribution (be that an intranet or the internet) does not fail. All of these major threats emphasise the importance of not rushing implementation and the necessity of thorough integrity testing. Also emphasised is the necessity of explicit disaster recovery plans, the most obvious of which would involve reserve power supplies, reserve machinery and regular data back-ups.

7.2.3.3 Assessment threats

The final, and most esoteric, threats concern the validity of assessment under the new technology model. These threats are esoteric because the assessment of achievement is so complex and poorly understood that it is very hard to predict whether changes in assessment practice risk changes in assessment product.

There are some obvious threats, though, the most obvious of which is that tension will arise between opportunities afforded by the marking technology and the nature of assessment valued by the educational community. A test that was marked entirely by computer would probably result in an external marking system that was significantly faster, cheaper and more reliable – but would it also be better? If computer marking proscribed constructed response formats then the UK educational community would be very reticent to agree. This kind of tension need not be quite so overt, yet still have a potentially negative impact. For example, as tests came to be designed with unskilled, or semi-skilled, markers in mind then questions might be written so as to discourage creative or idiosyncratic responses.

Less obvious threats might concern the way in which markers approached the task of marking using new technology. The crucial characteristic of the Pilot model was that pupils' responses were disembodied. That is, markers no longer marked a script with a name on the front indicating a specific pupil from a particular school (indeed, a local school, to whom a conventional marker would implicitly be accountable). The loss of feeling for real pupils from real schools could conceivably result in a subtle lowering of personal marking standards.

There are even potential threats from changes that increase the validity of marks arising from the external marking system. Whatever the genuine benefit from conventional borderline procedures, there is no doubt that they ultimately distort national curriculum test mark distributions. With the elimination of borderline checks the new distributions would differ significantly with implications for the maintenance of standards over time. Similarly, if conventional marking procedures happened to encourage markers to give pupils the benefit of the doubt (perhaps because they were seen as real pupils from real schools) then, in eliminating this bias, new technology marks might be more accurate. However, once again, if standards were to be maintained over time this would be a problem (as the new technology marks would tend to be lower even if the quality of the cohort was identical).

Allied to the concern that schools will not receive feedback from a specific marker is the concern that marking anomalies that affect specific schools may go undetected under the new technology model. Thus, if a particular teacher had taught pupils a particular way of responding to certain questions – and this response proved problematic for markers to relate to the mark scheme – then there would be the risk that some pupils would be rewarded while others would not. This would presumably not happen under the conventional model, as the marker would spot the anomaly, seek a ruling from the Chief Marker, and mark accordingly. Inconsistencies resulting from the new technology system would be irritating for teachers and might lead to requests for school reviews (in contrast to the earlier suggestion that new technology marking would eliminate R3 reviews). Similarly, the fact of marking scripts from the same school has, in the past, enabled markers to identify features that were indicative of malpractice – and this avenue for detection would no longer be available in a new technology system.

Developing the marking ambiguity concern, it should be noted that mark schemes are not set in tablets of stone once marking has begun and Chief Markers frequently relay new advice to markers during the marking period upon discovery of new ambiguities. It would be important to ensure not only that a mechanism still existed for ensuring this kind of communication, but also that mechanisms still existed for detecting these kinds of ambiguities. Once again, they are far more salient when they recur within scripts from a single school.

As mentioned in Section 2, there are potential threats to assessment that arise from the new marking apparatus itself. If, for example, different markers chose to mark handwriting at different magnifications (or if the same marker chose to apply the zoom for only certain pupils) this could have implications for the validity of the process. Judging handwriting at 100% magnification (i.e., as it would appear on paper) is very clearly not the same task as judging it at 400% magnification, and it would not be surprising if this added 'construct-irrelevant variance' to results that are not that reliable anyway (as the Pilot demonstrated).

7.3 Potential benefits, risks and costs associated with the web-based on-line marking model

While the previous discussion focused specifically upon the feasibility of scaling up to a centre-based model, many (if not most) of the points raised applied more generally to the feasibility of on-line marking, per se. Yet there are important differences between a centre-based and a web-based model. The following discussion will consider some of the

issues that would be of particular relevance to a national web-based on-line marking model. It is broken down according to the two main features that would set it apart from a centre-based model.

7.3.1 Web-based marking

There was a strong feeling from the stakeholders interviewed for Section 4 that a web-based model should be investigated further (and there was a suggestion that the final system might utilise both models). There would certainly be major benefits to be gained from the successful implementation of web-based marking, perhaps the most important being that markers would be very likely to find the prospect of marking at home very attractive. It has already been noted that a lack of participation from expert markers is a major threat to the success of a centre-based model.

Yet there are problems that would have to be investigated before taking the web-based option further, the most obvious of these concerning computer hardware and software. The idea of markers marking from home assumes that they would have access to hardware and software that could support the on-line marking process. But would markers be proscribed, or practically discouraged, from marking if they did not have such facilities? Whether it would be feasible to provide markers with the requisite technology during the marking period is an option that might be considered, but it sounds somewhat costly. Even markers who did have computer systems at home might face problems of compatibility or processing speed, particularly if they ran older machines. Indeed, it is yet to be established that the UK internet could effectively support on-line marking at an appropriate speed and cost.

Differences in hardware used by markers might even have implications for the payment model employed, as apparent productivity might actually depend more upon the efficiency of the PC on which the marking was conducted. There are also wider financial concerns to address, for example, if damage to personal machines occurred during the course of marking would the QCA be obliged to provide compensation?

Then there are security concerns. If markers were to work at public terminals rather than at home then the risk that confidential information could be accessed would need to be investigated. The storage of such information in browsers is a particular concern. Whether there may be similar risks associated with the theft of PCs should be considered. A potentially serious issue is whether the marking agency could verify the identities of those accessing the databases, particularly if training occurred before tests had gone live. Furthermore, with access via the internet, it would be essential that sufficient steps were

taken to prevent hacking or corruption by viruses. Hacking would be more of a threat the higher the stakes associated with the tests were.

An entirely distributed network of markers would also face threats associated with technical support. Would it be possible (within reasonable financial constraints) to support a very large network of markers working throughout the UK?

7.3.2 Exclusively web-based training and supervision

One of the major benefits of a web-based model would be realised if it proved possible to train and supervise markers entirely over the web. This would reduce the considerable costs presently associated with marker training. Furthermore, the level of automation implied by this model would ensure consistency and would free senior markers to devote more time to markers who were experiencing problems.

The threats associated with on-line training and supervision are less technical than those discussed above; they generally question whether it could be as effective as face-to-face approaches. It is fair to say that conventional training and supervision places a considerable emphasis upon discussion and social interaction. (The importance of discussion is explicitly stated in marker guides.) Yet discussion will not be possible in the same way via a web-based system. And even if interactive sessions could be organised, it is not clear that they would facilitate the same learning experience as face-to-face sessions do. The senior markers within the Pilot were not at all confident that on-line supervision could work. Moreover, there were concerns that the lack of personal contact would prevent senior markers from gaining insight into the personal skills of their marking teams, which is necessary for them to determine who might be appropriate for promotion to supervisory positions.

While a national web-based on-line marking model has much to recommend it, there are clearly major issues that would need to be addressed before scaling up could be considered. Of course, it should not be forgotten that it would be possible to develop any of a number of hybrid models, involving both centre-based and web-based marking. This might help to overcome some of the problems that might result through exclusive reliance upon one approach rather than another. For example, it might be possible to employ non-experts to work in marking centres, while experts marked from home.

7.4 Conclusions and recommendations

7.4.1 Conclusions

In response to the four over-arching objectives presented at the outset, the Evaluation team reached the following general conclusions:

7.4.1.1 Objective 1: to evaluate whether the NTP contractor managed successfully to implement the agreed procedures for 2001

NCS Pearson experienced two major implementation problems that compromised the extent to which conclusions could be drawn from the Pilot: they were unable to appoint a sufficient number of appropriately qualified expert markers; and they were unable to scan and mark the full complement of scripts. However, at least to some extent, these problems were beyond their control. They were somewhat more responsible for the decision to employ Netgrade software rather their full ePEN software, which also limited conclusions that could be drawn from the Pilot.

Overall, NCS Pearson were successful in implementing the agreed procedures for 2001 and demonstrated, at least on a small scale, that the QCA's ambitions for revamping the national curriculum external marking system are technically feasible.

7.4.1.2 Objectives 2&3: to evaluate whether the procedures implemented during 2001 were effective in delivering significant benefits without undue costs, and might be scaled-up successfully for all national curriculum tests

It will not be possible to predict confidently whether centre-based on-line marking is likely to improve the speed, accuracy, management, stakeholder satisfaction and financial viability of external marking until it is tested:

- with a high volume of scripts, delivered on time;
- with markers that are demonstrably of the conventional standard, but also familiar with new technology marking;
- with procedures that embrace the full range of conventional processes;
- with procedures that exploit the full potential of the new technology;

- having overcome technological teething problems;
- with pupils sampled from across the full ability range; and
- with results provided for every single pupil in the sample.

However, despite the inevitable limitations of the Pilot, in the opinion of the Evaluation team, a national centre-based on-line marking model would have the potential to improve the speed and accuracy of the external marking system, and thereby to improve stakeholder satisfaction, ***as long as a sufficient number of expert markers were prepared to embrace the culture change associated with marking in centres rather than at home.*** There are also reasons to believe that the management of the external marking system might be improved by the adoption of new technology, although this benefit could easily be eliminated by too many contractors and poor co-ordination between them. Finally, the Evaluation team is not convinced that a national centre-based on-line marking model would lead to significant financial savings. Indeed, it seems quite possible that it would increase costs. The benefits that arise would be in terms of faster turnaround, enhanced reliability and greater provision of information.

7.4.1.3 Objective 4: to consider whether revised procedures for future years might deliver significant benefits without undue costs

As above, it will not be possible to predict confidently whether a web-based system is likely to improve external marking until it is tested under realistic conditions. However, if it could be shown that the centre-based marking model had serious potential, then it would seem fair to conclude that the web-based model might extend the benefits to be reaped further still. The major advantage would seem to be flexibility for expert markers to work at home. To reduce the burdens upon expert markers – a precious and limited human resource – was one of the major goals of the new technology programme; and this can only be fully maximised through the provision of a web-based marking facility. Moreover, anticipated financial savings relating to travel and accommodation can only be realised through a web-based implementation.

Unfortunately, the web-based marking model brings additional threats – threats to the technical viability of the process (resulting from hardware and software obstacles) and to its quality (through a loss of face-to-face interaction). These may not be straightforward to overcome.

7.4.2 Recommendations

1. Further piloting for the new technology programme should be undertaken during 2002, although the system is not yet ready to be formally introduced for all or part of a national curriculum test where results stood. Further research into the feasibility of centre-based models, web-based models, and hybrid models should be conducted. However, it should not be forgotten that changes to conventional procedures – such as centre-based marking or the wider use of Optical Mark Reading – might also yield some of the desired benefits of a new technology system.

7.4.2.1 For immediate attention

2. Further consideration should be given to the fundamental aims of the new technology programme, and potentially competing goals should be explicitly weighted. As the saying goes: "faster, better, cheaper – pick any two." In particular, it may not be possible to ensure financial savings through the implementation of new technology. If this is not an acceptable consequence then it needs to be clarified at the outset. However, if it is an acceptable consequence, then at least rough parameters should be set for additional costs. Above all, it would seem essential that clear benefits be reaped in at least one of the five major areas – speed, accuracy, stakeholder satisfaction, management or finance – to avoid the charge of change for the sake of change.

3. The parameters of change need to be explicitly formulated. It may be that radical financial savings could be achieved, but only by marking tests exclusively by computer. If this is not acceptable then it needs to be made clear at the outset. This will involve explicit consideration of the extent to which new technology should be required to accommodate to the UK education culture, and the extent to which the UK education culture could, or should, be persuaded to embrace the new technology.

4. The parameters of risk tolerance need also to be explicitly formulated. The implementation of new technology will inevitably involve risk. To the extent that risks can be quantified, programme decision-makers need to be reassured that these are not too great (regardless of potential benefits).

7.4.2.2 Planning for future pilots

5. It needs to be decided whether the on-line marking model will aim for double marking of all expert items or single marking plus sampling.

6. Dependent on the decision between single and double marking for expert markers, models of training, standardisation, 'certification', supervision and monitoring will need to be designed. It will be important that these are tailored to the specific needs of the chosen approach and that they are effectively piloted. Software features should be driven by the model of quality assurance decided upon (rather than the choice of quality assurance model being driven by software features provided). Particularly if single marking is chosen, it will be important to ensure that markers have greater flexibility to return to items that they are unsure about, and that the potential for accidental mis-selection of mark options is reduced.

7. Formal plans should be developed to define data needs within the new technology system, and to specify how data will be produced and used.

8. Future contractors should provide a solution to the problem of pupils' responses being located beyond the clip image areas.

9. Future contractors should ensure that the new technology systems are sufficiently flexible to cope with special arrangements for assessment.

7.4.2.3 Programme management

10. During all key decision-making stages, consultation and negotiation with markers should be undertaken. Without the support of expert markers the new technology programme is unlikely to succeed (unless the role of the expert marker is to be reduced very significantly).

11. Where feasible, medium-term as well as short-term plans ought to be developed and disseminated. Short-term contracts do not necessarily support effective development.

12. Debates over test format changes need to occur as early as possible, as national curriculum test development occurs two years in advance of testing.

13. Computer programs underlying future systems should be scrutinised by experts from a range of fields to ensure that they effectively achieve what they are intended to. Experts in computing, psychometrics and quality assurance will need to be involved, as well as project managers and senior markers.

7.4.2.4 Specific research requirements

14. Further research should be conducted into the relationship between marks from conventional and new technology systems. A number of important differences were observed during the Pilot (particularly with respect to mean mark differences) and it will be essential to determine whether these effects are replicable. Future research should explore direct comparisons between conventional and new technology marks (preferably against the 'gold standard' of Chief Markers). It should also explore further comparisons within each of the systems (such that data on marking reliability for handwriting can be determined for conventional, as well as for on-line, markers).

15. Research should be conducted into reasons for adjudication and steps that can be taken to reduce adjudication demands. Although the Pilot supported the use of clerical and response selection markers for marking simpler questions, any steps that could be taken to reduce the adjudication load would be likely to reap significant benefits.

16. Research should be conducted into models of work allocation and marker payment.

Section 8 References

MURPHY, R.J.L. (1978). 'Reliability of marking in eight GCE examinations', *British Journal of Educational Psychology*, **48**, 2, 196–200.

MURPHY, R.J.L. (1982). 'A further report of investigations into the reliability of marking of GCE examinations', *British Journal of Educational Psychology*, **52**, 1, 58–63.

NEWTON, P.E. (1996). 'The reliability of marking of General Certificate of Secondary Education scripts: mathematics and English', *British Educational Research Journal*, **22**, 4, 405–20.

WHETTON, C., TWIST, L. and SAVAGE, R. (1998). Further Development of Mark Schemes for Writing for Key Stage 2 English. Unpublished report.

NFER HEAD OFFICE
National Foundation
for Educational Research
The Mere
Upton Park
Slough
Berks SL1 2DQ.
Tel: 01753 574123
Fax: 01753 691632
E-mail: enquiries@nfer.ac.uk
Web site:http: //www.nfer.ac.uk

NFER WELSH OFFICE
Chestnut House
Tawe Business Village
Phoenix Way
Enterprise Park
Swansea
SA7 9LA.
Tel: 01792 459800
Fax: 01792 797815
E-mail: scyanfer@abertawe.u-net.com

NFER NORTHERN OFFICE
Genesis 4
York Science Park
University Road
Heslington
York
YO10 5DQ.
Tel: 01904 433435
Fax: 01904 433436
E-mail: jbh3@york.ac.uk