National Foundation
for Educational Research

An Evaluation of the 2002 New Technologies Pilot

A report for the Qualifications and Curriculum Authority

Catherine Kirkup, Paul Newton, Ewan Adams, Nicola Page and Chris Whetton

FINAL REPORT

29 November 2002

Acknowledgements

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

NFER Project Team

Project Director:	Chris Whetton
Project Leader:	Paul Newton/ Catherine Kirkup
Researchers:	Ewan Adams, Nicola Page, Bethan Burge
Statisticians:	Rachel Dingle, Emma Scott

Additional NFER staff

Joan Howell, Humaira Ishaq, Margaret Parfitt

QCA Project Team

Cleo Nicolaidou-Wright, Helen Patrick, James Butler and colleagues

NCS Pearson Project Team

Kris Knowles, Brian Speed and colleagues.

The Marking Teams

David Spratt, Doug Grieves, Margaret Cooke and all members of the key stage 3 mathematics marking teams.

Contents

		Page
Section 1	**Introduction**	1
1.1	The 2001 New Technologies Pilot	1
1.2	The 2002 New Technologies Pilot	2
1.3	The 2002 Evaluation – aims and objectives	4
1.4	The 2002 Evaluation – methodology	12
1.5	The evaluation report	16
Section 2	**The implementation of agreed procedures**	**17**
2.1	Methodology	17
2.2	The functioning of the pilot	19
2.3	Study 6 – frequency of pupils responses located beyond the clip image area	56
2.4	Summary of key issues arising	66
Section 3	**Questionnaire feedback from e-markers**	**68**
3.1	Introduction	68
3.2	Methodology	68
3.3	Training	69
3.4	Software/technical issues	70
3.5	Technical support	72
3.6	Supervision	72
3.7	Expectations and experience of e-marking	73
3.8	Views on scaling-up the system	75
3.9	Other comments	76
3.10	Summary	77

Section 4	**A statistical analysis of the pilot**	**79**
4.1	Analysis of management data	80
4.2	Analysis of measurement data	94
4.3	Study 5 – a controlled evaluation of marking reliability	111
4.4	Summary	119
Section 5	**An evaluation of the 2002 New Technologies Pilot**	**123**
5.1	Specific objectives for 2002	123
5.2	Conclusions	131
5.3	Recommendations	136
Appendices	**See separate document**	

Section 1 Introduction

The 2002 New Technologies Pilot is an extension of work sponsored by the QCA, delivered by NCS Pearson and evaluated by the NFER, during 2001. The 2001 pilot was the first in a series of projects that would explore the feasibility of implementation of an electronic system for the marking of national curriculum test scripts and subsequent data collection.

1.1 The 2001 New Technologies Pilot

The defining feature of the 2001 pilot was the scanning of electronic images of students' scripts into a central computer. As part of this process, responses to individual questions from each script were scanned into a database as separate images. These discrete item image 'clips' were then distributed to teams of 'e-markers' who marked them on-screen using software delivered over an intranet. This item distribution process enabled different types of question to be marked by different types of marker, and allowed markers to specialise by marking responses for only one (or a small number of) question(s) during any particular marking session.

Questions from the tests under investigation – the year 7 progress tests – were divided between three categories, relating to the level of skill required to mark them:

1. unskilled (referred to as 'data entry' questions);
2. semi-skilled (referred to as 'clerical' questions);
3. skilled (referred to as 'expert' questions).

Within the new technology system, only expert questions were marked by traditional markers (i.e., those with teaching experience). Temporary staff without subject matter knowledge were employed to mark clerical and data entry questions.

The key components of the 2001 electronic marking model were as follows:

- advance script personalisation and electronic tracking through all processing stages;
- script batching, guillotining (to remove spines prior to scanning) and script 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 student performance for schools, prepared electronically and accompanied by electronic script images.

1.2 The 2002 New Technologies Pilot

The intention of the 2002 pilot was to extend the methodology developed during 2001, to explore further the potential of new technology systems and processes.

1.2.1 The structure of the 2002 pilot

At the close of the 2001 pilot, one of the most important outstanding questions was whether the scanning technology could deal with the volume of scripts that would be involved if the process were scaled-up to a national level. Although it would not have been appropriate to use an entire cohort for the 2002 pilot, the sample of scripts to be scanned was increased significantly. For 2002, it was intended to scan the entire work of around 20,000 students (although the eventual number was closer to 15,000).

It was also considered important to extend the pilot to a test that was taken by students across a full ability range (cf. progress tests, which are restricted to lower ability students). The decision to focus upon key stage 3 mathematics for the 2002 pilot meant that the full ability range would be sampled and also that the new technology would be employed with an older group of students. Table 1.1 illustrates the nine possible combinations of papers that can be taken for key stage 3 maths.

The third crucial characteristic of the 2002 pilot was to be its extension to web-based expert markers. A particularly attractive feature of the new technology system is that it

allows e-markers to work from home (via the internet) rather than from a marking centre (via an intranet). Only the latter was piloted during 2001 and the inclusion of a 'web-based' marking model, in addition to a 'centre-based' marking model was an important advance for 2002.

Table 1.1 The nine possible combinations of papers for key stage 3 mathematics

Combination	Paper 1 (no calc.)	Paper 2 (calc)	Mental Arithmetic	Extension
1	tier 3 to 5	tier 3 to 5	C	n/a
2	tier 4 to 6	tier 4 to 6	A	n/a
3	tier 4 to 6	tier 4 to 6	B	n/a
4	tier 5 to 7	tier 5 to 7	A	n/a
5	tier 5 to 7	tier 5 to 7	B	n/a
6	tier 6 to 8	tier 6 to 8	A	n/a
7	tier 6 to 8	tier 6 to 8	B	n/a
8	tier 6 to 8	tier 6 to 8	A	Extension
9	tier 6 to 8	tier 6 to 8	B	Extension

The major stages of the 2002 pilot were rolled-out to the schedule presented in Table 1.2.

Table 1.2 Major stages of the 2002 pilot

Time	Action
Jan – Apr 2002	Design, print and despatch modified test papers
May 2002	Receive completed scripts directly from schools Batch, guillotine and scan scripts Collate, re-batch and despatch scripts to conventional markers
July 2002	Receive sub-sample of conventionally marked scripts from schools Input data from conventional markers Centre-based e-marking of data entry and clerical items Web-based e-marking of expert items
Aug 2002	Centre-based e-marking of expert items
Sept 2002	Preparation of data from pilot for evaluation team Preparation of reports on student performance for schools

1.2.2 A summary of structural and procedural changes

In addition to the three major changes for 2002 outlined above, there were a number of other important differences. The principal changes for 2002 are summarised below:

- the scanning of a larger number of scripts;
- the capture of neither full-page images nor images from scripts that failed to scan;
- a focus upon key stage 3 maths tests (rather than the year 7 progress tests in maths and English);
- the parallel operation of centre-based and web-based marking models;
- advance script personalisation at the school-level, but not at the student-level (i.e., unique script identifiers but not pre-printed names);
- electronic response capture prior to conventional marking (such that markers' comments could not be captured accidentally);
- a delay between scanning and e-marking (to enable the appointment of experienced conventional markers to expert e-marking posts);
- a 100% double-marking model for expert items (as well as for data entry and clerical items);
- the use of ePEN e-marking software (rather than Netgrade);
- improved on-line mark scheme functionality.

1.3 The 2002 Evaluation – aims and objectives

In essence, the aim of the 2002 evaluation was to provide an independent judgement of whether new technology marking has the potential to revolutionise the UK's external marking and data collection systems for national curriculum tests. The breadth of this remit reflected the importance and scope of the project.

1.3.1 Four general objectives for 2002

To make the evaluation task tractable, this high-level aim was translated into four general objectives:

1. to evaluate whether the NTP contractor managed successfully to implement the agreed procedures for 2002;

2. to evaluate whether the procedures implemented during 2002 were effective in delivering significant benefits without undue costs;[1]

3. to consider whether the procedures implemented during 2002 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.

In fact, these were also the objectives that guided the 2001 evaluation.

1.3.2 Seven specific objectives for 2002

As the project developed, the QCA also highlighted a range of specific questions for which answers were required; these related to particular design features of the 2002 pilot. As such, seven specific objectives were proposed as principal foci for the 2002 evaluation:

1. to compare conventional and new technology marking standards;

2. to highlight significant marker, and marker training, issues arising from the 2002 e-marking models;

3. to compare strengths and weaknesses between centre-based and web-based e-marking models;

4. to explore the merits of double-marking;

5. to identify process scale-up issues;

6. to determine whether management information systems functioned optimally;

7. to evaluate the effectiveness of system improvements for 2002.

Each of these will be discussed below, in turn.

---

[1] Costs are to be understood as any negative consequence and not simply in financial terms.

1.3.2.1 Conventional versus new technology marking standards

The 2001 evaluation identified a small, though significant, reduction in mean marks awarded to students through the new technology process. It was not at all clear what was responsible for this. One suggestion was that conventional markers may have been more liable to give the 'benefit of the doubt' to students who were seen as individuals rather than as de-personalised item clips. On the other hand, higher conventional marks were even observed in relation to data entry questions – questions that ought to have been unambiguously right or wrong – and this casts an element of doubt on the 'benefit of the doubt' hypothesis. An alternative explanation is that new technology marks may have been lost by students who recorded valid responses in areas not visible to e-markers (i.e., entirely beyond the clip image areas).

It was considered crucial that this finding be explored further. As such, it was decided that the 2002 pilot should replicate the comparison of conventional and new technology marks.

If the mean mark differences were not replicated during 2002 then this would raise a question mark against the significance of the effect observed during 2001 (it may have occurred 'by chance' or may have been caused by uncontrolled process or design factors that went unnoticed). On the other hand, if the differences were replicated during 2002, then this would have significant implications for the wider roll-out.

1.3.2.2 Marker, and marker training, issues

The nature of interaction between markers and new technology processes is clearly central to the success of the enterprise. The evaluation was therefore designed to consider this interaction from a variety of perspectives:

- cognitive (knowledge and understanding of systems and processes);
- conative (intentions and aspirations for systems and processes);
- affective (feelings and attitudes towards systems and processes).

These were to be explored through a variety of contexts, in particular:

- marker recruitment and appointment;
- marker training and qualification;

- marking praxis;
- marker supervision and monitoring.

1.3.2.3 Centre- versus web-based marking processes

A major innovation of the 2002 pilot was the incorporation of web-based marking. A group of expert markers was appointed to work from home, at their own PCs, drawing image clips from a central database in Hellaby via the internet.

The evaluation aimed to consider issues such as those raised in 1.3.2.2 (above), for centre- and web-based expert markers respectively, to support a comparative analysis of their relative strengths and weaknesses. It aimed also to consider further issues related to support requirements and technological infrastructure.

1.3.2.4 Merits of double-marking

Another innovation was the decision to employ a 100% double-marking model for all scripts and all marker types. This contrasted with the 2001 pilot which (for expert markers) had an ambiguous marking model, but one that was ultimately based upon single-marking.

The double-marking model is more familiar to NCS Pearson and ePEN functionality has ultimately been designed with the double-marking model in mind. It was assumed that 100% double-marking would allow NCS Pearson to implement procedures for training, qualification, calibration, monitoring and validation that were said to have been demonstrated (in the USA) to be effective.

The evaluation proposed to consider how effectively these procedures translated to the UK context, focussing upon the roles and responsibilities of the e-markers and, in particular, their supervisors (and the nature of demands made upon them).

Double-marking will, inevitably, result in improved marking reliability. However, it has obvious costs in terms of the increased number of markers required, the increased financial expense of marking, and the decrease in marking speed. The evaluation aimed to collect data to support an empirical estimation of these impacts (although financial costs were not to be evaluated directly).

Importantly, it should be recognised that the question that ultimately needs to be answered is not: "what is the optimum proportion of items to double-mark?" The principal question is: "should a single- or a (100%) double-marking model be employed?" Any re-marking

that occurs within the single-marking model occurs ultimately for monitoring purposes. The point of re-marking within the single-marking model is *not* to improve marking reliability by a factor of X or Y (determined by the proportion of items re-marked). Re-marking is employed essentially to establish whether or not markers are marking to an acceptable standard – and the frequency and timing of re-marking, within the single-marking model, should be decided purely in terms of its quality assurance impact.[2]

To clarify this point, the stakes associated with marking error within a single-marking model are very high since (assuming that fewer than 50% of items are re-marked) most items will be marked only once. This means that a marking agency would have to be very confident – at a very early stage – that all markers were marking at a standard that was as high as could be expected. In contrast, the stakes associated with marking error within a double-marking model are less high, as all items are marked twice, and the vast majority of problems that occur in the early marking stages – as markers develop confidence and experience – would automatically be identified for adjudication. This difference in stakes leads to the question of whether training, qualification and supervision models ought also to differ. An implication might be that the double-marking model could risk a less intensive burden of early training and support (in favour of speed of marking, say) while the single-marking model could not risk this.[3]

Once again, the ultimate decision is whether to adopt a single-marking model or whether to adopt a double-marking model. Unfortunately, although the pilot was designed to provide information relevant to the costs associated with double-marking (re speed, human resources and finance), it was not designed to provide comparable information on the costs associated with an alternative single-marking model. (No alternative single-marking models for the UK context have yet been specified.) Nor was the evaluation designed to deliver recommendations such as the optimum proportion of items to re-mark within an alternative single-marking model (as evidence necessary to support this kind of

---

[2] Note that the idea of adjudicating re-marked items within a single-marking model is, in a sense, something of a red-herring (i.e., you monitor primarily for quality assurance of the process – to make sure that markers are marking effectively – and only secondarily to correct incorrect marks that you happen to notice along the way).

[3] As it happens, the single-marking models that are used in the USA by NCS Pearson are essentially the same as those used for double-marking. However, it should not be assumed that this will also be true for the UK.

inference would not emerge from a pilot that was exclusively concerned with double-marking).[4]

1.3.2.5 Process scale-up issues

The principal scale-up issue addressed by the 2002 pilot was to be the scanning of large numbers of students' scripts. Unfortunately, the approach of the pilot was limited, as it was explicitly decided not to scan problematic scripts, i.e., the very scripts that cause scanning delay. One reason for this was the suggestion that more effectively designed papers might not suffer from the same scanning problems: having to scan all problematic scripts might therefore provide an unrealistic estimate of future scanning rates; moreover, it might compromise the 2002 effort to provide information on optimal scanning rates.

Full sets of scripts for around 20,000 students were to be scanned during cycle 1. The figure of 20,000 students was intended to translate into just over 60,000 student scripts.[5]

In relation to the scanning scale-up, the evaluation intended to consider:

- how effectively NCS Pearson managed the transmission of scripts through the various stages of the scanning cycle;
- the optimum scanning speed that NCS Pearson were able to achieve through rejection of problematic scripts;
- the frequency with which problematic scripts were encountered and the reasons for their rejection.

Scaling-up would also be considered more generally in relation to the kinds of issues that were discussed in the 2001 pilot, for example:

- hardware and software issues (especially in relation to web-based marking);
- marking praxis;

---

[4] Note that reliability data will still provide evidence with which to estimate the degree to which re-marking reduces the marking error rate.

[5] The original estimate of 20,000 students was subsequently reduced to 15,000; likewise, the estimate of 60,000 was reduced to somewhere in excess of 45,000.

- training and supervision models;
- etc..

1.3.2.6 Management information systems

As discussed within 1.3.2.4 (above), the double-marking model meant that different types of management information were to be collected for different purposes and used in different ways (that is, when compared with the 2001 pilot).

The NFER proposed to evaluate the information generated by NCS Pearson and the manner in which it was used, particularly in relation to the supervision of markers.

1.3.2.7 System improvements

The last of the QCA's specific concerns was to evaluate the effectiveness of procedural and systemic changes that had been implemented for the 2002 pilot. These factors included:

- the use of ePEN rather than Netgrade software;
- the appointment of experienced markers to expert e-marking positions;
- the improvement of on-line mark schemes.

Other changes to systems, processes and procedures would be identified and explored as the pilot progressed (primarily through Study 2 of the evaluation).

1.3.3 Guiding principle of the 2002 evaluation

The guiding principle of the evaluation was to determine the strength of threat, posed by various factors, to the scale-up of the new technology system. This meant distinguishing between factors with potentially major impacts upon the feasibility of scale-up and those likely to have no more than moderate or trivial impacts. In short, the aim was to identify the principal obstacles to successful scale-up and, where possible, to make suggestions as to how these might be overcome. The evaluation took account of the following principal categories of risk, as defined in the New Technologies Pilot Project Initiation Document (PID version 0.3, 15 April 2002):

1. risks associated with increasing the volume and scaling of scanning;
2. risks linked to the technology;

3. risks associated with the cost of the redesigned process;

4. risks resulting from the impact of change.

The evaluation was framed by the QCA's 'ultimate vision' for the implementation of new technology (by 2007). As discussed in the *Vision and Blueprint* document for England and in the NTP 2002 PID, this aspiration was characterised by the following elements:

1. the EMDC (external marking and data collection) process is streamlined such that it reduces the administrative burden on schools;

2. the EMDC process is streamlined such that it is able to provide more valuable outcomes from the tests for schools;

3. the EMDC process is streamlined such that it provides more accurate data for use by DfES, OFSTED and LEAs;

4. process improvements will increase efficiency, drive down cost, reduce risk, require fewer specialised services and lead to a less complicated and more widely understood system;

5. the marking process is streamlined such that it removes the administrative burden on markers;

6. the marking process is streamlined to make more efficient use of marker expertise;

7. there is better and more effective preparation and supervision of markers.

Amongst these elements, the fourth was one of the most crucial. The desire to increase efficiency and to reduce risk were the principal drivers here. Ultimately, the intention of the new technology programme is to improve value for money, rather than to reduce financial costs, *per se*. A particular aspiration for the EMDC process is that a structural simplification will open the various aspects of external marking and data collection to a greater number of potential suppliers – enhancing the degree of competition within the sector. The breakdown of EMDC into structurally independent modules (e.g., test design, scanning, marking, review, etc.) is intended to facilitate this.

The evaluation was also guided by two high level outcomes that defined success criteria for the project (again from the NTP 2002 PID):

1. on project completion QCA must be in a position to make a decision on whether to proceed with electronic marking of one subject, nation-wide, in 2004/5 (i.e., the pilot must provide sufficient information to support a valid decision);

2. the project must inform the EMDC procurement exercise for 2005-2007 in relation to the capability requirement of prospective contractors (i.e., the pilot must provide sufficient information to support the development of a valid specification for a revised EMDC process).

1.4 The 2002 Evaluation – methodology

The final integrated proposal from the NFER specified that the general and specific objectives of the evaluation would be met through a series of seven inter-linked studies. These are summarised below.

1.4.1 Study 1 – The development of a data collection and evaluation framework

Study 1 encapsulated the research and development that was required for the first stage of the evaluation project. This involved detailed discussions of the proposed methodology (between NFER, QCA and NCS Pearson) as well as an analysis of supporting documentation. This groundwork ensured that NFER staff were fully appraised of the scope of the pilot and of the practical and technical constraints within which it was required to function. It supported the developmental component of Study 1, during which NFER produced a framework for the collection and analysis of data from the pilot (Newton and Whetton, 2002).

1.4.2 Study 2 – The implementation of agreed procedures

Study 2 was intended to answer questions related to the implementation of agreed procedures by NCS Pearson. It involved two main research components: an analysis of documentation; and an observation of the pilot.

An analysis of documentation

The analysis of documentation produced in service of the pilot supported three main functions, related to three different kinds of document:

1. **specification documents** (e.g., exemplar DWS reports, e-marker recruitment plans, etc.) – to ensure that NFER staff were fully appraised of the way in which the system was intended to function during 2002;

2. **support and training documents** (e.g., training materials, User Guides, desk instructions, etc.) – to clarify intended procedures and to ground an evaluation of support and training functions;

3. **outcome documents and reports** (e.g., exemplar test papers, issue logs, etc.) – to ground an evaluation of the effectiveness of various aspects of the system.

An observation of the pilot

The observation of the pilot involved collecting evidence on the effectiveness of pilot procedures via:

1. **informal verbal feedback** from NCS Pearson project managers/supervisors;

2. **informal verbal feedback** from QCA project managers;

3. **informal verbal feedback** from e-marking supervisors and e-markers;

4. **direct observation** of the e-marking process.

An NFER observer was present at the marking centre for most of the duration of the main e-marking period. This included an observation of training delivered for e-markers and an observation of e-marking and supervisory functions. More attention was focused upon clerical e-markers and (particularly) centre-based expert e-markers, although there was some direct observation of data entry. Feedback from web-based expert e-markers was achieved by telephone.

1.4.3 Study 3 – Questionnaire feedback from e-markers

A questionnaire was despatched to all web- and centre-based expert e-markers at the end of the marking period. Questions focused upon their perceptions of the strengths and weaknesses of the e-marking system as they experienced it during the pilot. The questionnaires explored e-markers' prior perceptions of new technology marking, whether they felt more or less positively disposed to the idea after having experienced it, and whether they felt that the new technology promoted high standards of marking.

1.4.4 Study 4 – A statistical analysis of the pilot

A major component of the 2002 evaluation was a statistical analysis of data arising. These tended to fall into one of two categories: management data (quantitative information on the speed and accuracy of procedures for the processing of scripts through

the electronic marking system); and measurement data (quantitative information on the accuracy of results arising from the electronic marking of scripts). More specifically, these included the following.

Management data:

- time taken to process scripts through guillotining and scanning stages;
- time taken to mark individual questions;
- supervisory and adjudication demands upon senior markers;
- prevalence of various processing errors.

Measurement data:

- paper- and subject-level correlation between conventional marks and e-marks;
- paper- and subject-level absolute mark differences between conventional marks and e-marks;
- question-level concordance statistics;
- subject-level concordance statistics.

1.4.5 Study 5 – A controlled evaluation of marking reliability

As part of Study 4, important data were collected on the reliability of marking of scripts. At an overall level, it was possible to compare marks that arose from conventional marking with those that arose from new technology marking. However, it was not possible to make direct comparisons because there was no control over which conventional markers had generated the marks and because the e-markers marked at item level rather than at script level.

For Study 5, a more controlled evaluation was achieved using an experimental manipulation in which four of the centre-based e-markers:

1. conventionally marked a set of 100 scripts (in addition to their basic conventional allocation);
2. electronically marked the same set of 100 scripts.

This study enabled production of the following statistics:

- between-marker reliability data for conventional marking;
- between-marker reliability data for new technology marking;
- within-marker reliability data for conventional versus new technology marking.

Neither analyses 1 nor 3 have been produced previously, nor has 2 been produced under such controlled circumstances.

1.4.6 Study 6 – Beyond the clip image area

The size of the item image clip remained an issue for 2002, just as it was during 2001. Indeed, as there was no scanning of full-page images during 2002, any problem arising from responses (or working) being located beyond the clip image areas was potentially more serious.

The investigation revolved around a manual scrutiny of full-page images of students' scripts. (NCS Pearson scanned a special sample of 350 scripts for this purpose.) These were inspected using the NCS Pearson Image Viewer with a clip image area electronic template overlay. The research was based in Slough and item-images were viewed remotely via the internet.

Only one question was investigated: does any part of a student's response to each item lie beyond the clip image area? This yielded information on the frequency of problematic responses. Data were presented separately by tier and by paper, giving an indication of which items were the most problematic.

1.4.7 Study 7 – Scanning exceptions

The final study was a collaboration between NCS Pearson and NFER to capture information on the frequency and nature of scanning exceptions – scripts that did not get scanned. In fact, these were divided into two categories:

1. exceptions (instances of regular scripts being rejected by the scanner, e.g., because students had scribbled on the timing marks);
2. attachments (instances of scripts being rejected due to the addition of extra sheets or notes).

One form was completed by scanning staff for each exception/attachment. Data from this exercise fed into the report on Study 4.

1.5 The evaluation report

The following sections present the results of the seven studies outlined above. The major studies form the basis of the major sections: Study 2 (Section 2); Study 3 (Section 3); and Study 4 (Section 4). Results from the other studies have been integrated within Sections 2 to 4 as appropriate. Section 5 presents an overall discussion of the evaluation objectives in light of the studies' results.

2 The implementation of agreed procedures

The principle aim of Section 2 is to achieve the first of the four general objectives specified in Section 1:

1. to evaluate whether the NTP contractor managed successfully to implement the agreed procedures for 2002.

This will be achieved using evidence gathered from across Studies 2 to 7, but primarily focusing upon findings from Study 2: The implementation of agreed procedures. This section outlines the methodology employed within Study 2 to determine the effectiveness or otherwise of the procedures implemented by NCS Pearson. It then considers, in turn, the major procedures within the pilot, giving a description of how each one was intended to function, according to the available documentation, and providing evidence from informal discussions and observations relating to these procedures. At the end of this section, strengths and weaknesses are summarised in order to make suggestions as to the extent to which such procedures might be effective on a much larger scale and to inform decisions about future development. Quantitative data relating to these procedures are reported in Section 4.

As will be described in more detail below, not all of the procedures were observed directly but depended on an analysis of documentation and discussions with NCS Pearson personnel. Thus, while an attempt has been made to discuss all the significant elements of the pilot, it is possible that some areas were not examined in sufficient depth. However, it is hoped that the major issues of relevance to this year's evaluation have been considered.

2.1 Methodology

The methodology involved three main components; a documentary review, participant observation of the training process and an observation of the marking process.

Key documents and specifications were submitted to the evaluation team by NCS Pearson in advance of the observational phase. These included technical specifications, exemplar materials, training documents, procedural guides and error/issue logs. In order to evaluate those procedures not observed directly, a review of this documentation was carried out, where necessary supplemented by discussion with NCS Pearson.

The role of the evaluation team and the purpose of the interviews (and, for centre-based markers, the observation) were explained to the markers in advance, by letter for the home-based markers and during training for the centre-based markers. They were also

given an indication of the broad issues that the researchers would be considering. Members of the evaluation team attended the training sessions for clerical markers, centre-based supervisors and centre-based markers. Participant observation allowed researchers to evaluate the training process as well as being appraised of the intended marking procedures of the pilot.

The observation of the marking process comprised three elements; *event shadowing*, *casual discussion* and *informal interviews*. *Event shadowing* involved simply circulating between the markers and noting events as they happened, for example technical problems with the software or network or discussions about marking issues between markers, or between markers and their supervisor. The aim was to be as unobtrusive as possible, not impeding or disrupting the marking in any way but using these 'natural' disruptions to solicit information about problems or concerns. Information was also obtained informally during *casual discussions* that arose naturally during any shared breaks and markers were encouraged to bring any significant issues to the attention of the NFER researchers during such breaks. The evaluation team also tried to capture the context in which marking was occurring, for example the organisation of the marking, the working environment and the general atmosphere of the marking centre.

Given that there were very few periods in which markers were unable to mark due to technical problems, the most fruitful source of information was provided by the *informal interviews*. These short, semi-structured interviews were carried out with both home-based and centre-based markers and supervisors during the period in which they were contracted to work. (Interviews with home-based markers were carried out by telephone, at pre-arranged times to suit the marker.) In the Framework document, the original intention was to enlist a number of representatives from each of the marking groups and approach them at regular intervals for informal feedback on the e-marking process. However, it was decided prior to the e-marking period that it would be more useful to involve all participants in the evaluation process in order to obtain views from as wide a range of markers as possible. Each marker and supervisor was therefore interviewed on at least one occasion (in most cases twice) and asked to comment on the strengths and weaknesses of the system as they experienced it. A number of issues and questions had been identified as potentially useful for guiding the observation elements and the interviews. Researchers therefore used a series of semi-structured prompts to ensure that similar topics were addressed in all of the markers' interviews. However, care was taken to ensure that markers did not feel constrained by this structure as unsolicited comments and questions posed by the markers to the researchers could in some instances reveal useful additional information. Home-based markers were asked some additional questions about the installation of the software, their operating system and internet

connection and the technical support they had received. A different set of prompts was used for the interviews with the supervisors. Copies of these prompts are presented in Appendix 1.

Notes from the informal interviews were recorded on individual pro-formas as a formal record of each interview but also to facilitate the collation of information across themes and markers.

2.2 The functioning of the pilot

2.2.1 Answer book modification and printing

To enable effective scanning of pupil scripts, it was necessary for NCS Pearson to make minor amendments to standard key stage 3 mathematics test papers, adding scanner 'timing track' marks and page identity marks. A bar code was also to be printed onto the cover of each test paper to allow individual paper tracking. NCS Pearson managed the print production process through its partner organisation Lonsdale.

Electronic copies of the standard test papers were provided by QCA to NCS Pearson and although there was a minor problem with incompatible fonts, design issues generally went well. Unlike the 2001 pilot, the 2002 scripts were not personalised at the individual pupil level. Print production figures were based on figures supplied by QCA, based on 2001 school and pupil registration information. Unfortunately, the actual demand from the participating schools was higher than expected, resulting in insufficient materials being printed for all 127 schools. NCS Pearson agreed with QCA to sample from these schools a total of 84 schools to meet the requirement for 20,000 pupils. Also, the barcode file had to be redone due to the recommended volume figures being underestimates.

Paper and print quality audits were carried out by NCS Pearson. A two per cent sample of all materials from the printers was subjected to a quality check. One such check of printed test papers revealed that an incorrect cover had been attached to one paper. Although this was within the agreed tolerances, NCS Pearson carried out a 100 per cent check to ensure a correct test body/cover match. This resulted in a small delay of a couple of days. Additionally, some papers had to be returned to Longsdale after a scanning test. An order for fulfilment materials to EC Logistics[6] was misinterpreted, resulting in a delivery of 924 header cards instead of 924 notes for teachers. These and other minor

[6] EC Logistics are a QCA sub-contractor with overall responsibility for printing and distributing test materials to schools.

difficulties were documented in an 'issues log' by NCS Pearson as part of their quality control procedures.

Generally, the difficulties faced were of the type that might be encountered with conventional test papers and were not excessive. The checks and procedures put in place by NCS Pearson were sufficient to overcome these difficulties.

2.2.2 Despatch of test materials to schools and test administration

The tracking of the test materials for the New Technologies Pilot 2002 was carried out by means of the Distributed Workflow System (DWS). Testing of the system for processing the key stage 3 mathematics test papers was carried out in March 2002. This included testing the systems for data initialisation, despatch to schools, batching, scanning and despatch to conventional markers.

Papers were despatched and individually tracked to participating schools. This process was managed through the DWS Mail Manager program. Prior to despatch, quality assurance checks were carried out for a sample of schools to check the quantity, type and contents of the packs for teachers and pupils and to ensure that all personalised documents had been bar-coded by the DWS MailMan program. The precise despatch content, time of despatch and despatch administrator were logged for each script of each batch. A help desk was established to deal with queries and requests for additional materials. Despite such quality assurance procedures, one school received a box of test papers that appeared to have 10 papers missing. The papers were later found, wrongly labelled in another box, but in the meantime the school had photocopied the missing test papers so that pupils could sit the test. Also, at least one school wanting extra materials contacted QCA directly rather than NCS Pearson and therefore received non-scannable versions. Such incidents were sources of frustration for participants. In any scaled-up operation it is essential that the packing of materials be very strictly controlled and that there is communication between all the relevant stakeholders to ensure participating schools receive the relevant materials.

Entries in the issues log indicate that locking up materials each evening was time consuming and that ensuring sufficient secure storage was occasionally problematic. In any future scaled-up procedure, consideration must be given to the massive increase in the volume of paper that would be involved and the resultant issues of security, storage and transportation.

After the key stage 3 tests had taken place, participating schools were requested to return scripts to NCS Pearson rather than to markers. This amounted to approximately 20,000

pupils and 60,000 pupil test papers. Despite these instructions, several schools sent their papers directly to AQA, resulting in their materials having to be excluded from the pilot. Only 76 of the 84 schools returned materials to NCS Pearson but it was agreed with QCA that there would be no chasing of non-returned scripts.

Some schools supplied their scripts alphabetically by pupil by tier whereas others supplied them alphabetically by whole pupil list. Due to the rapid turnaround required at NCS Pearson it was agreed with QCA not to re-sort the scripts as per the order on the marksheets. This subsequently caused problems when the scripts were despatched to conventional markers.

2.2.3 Test processing cycles 1 and 2

From the receipt of the test booklets, a target of 48 hours turnaround was set for NCS Pearson to batch, slit, scan, re-collate and package the materials, ready for despatch to the appropriate AQA appointed markers. At each stage, the processing of scripts was managed through the DWS Batch Builder and Batch Tracker programs.

On receipt, the precise receipt content, time of receipt and receipt administrator were automatically registered for each script of each batch. Once the receipt of the completed scripts had been logged in, test papers were batched in quantities of 50 according to type, with a header sheet attached to each batch to track the materials through the system. Spines were removed by slitting/guillotining to form sheets suitable for scanning. The documents were then passed through a high-speed document scanner to capture the image clips of pupil response areas. These clips were archived for later use in Cycle 2, the electronic marking element of the project. No scanning of attachment sheets was carried out.

The main purpose of the scanning pilot was to evaluate to what extent the amount of time required to scan scripts has a negative impact on the time available for marking. It had already been agreed that documents that could not be scanned through the high-speed scanner and any attachment documents would not be processed during this pilot. Any scanning exceptions, attachments or additional pages were recorded on a Scanning Exceptions/Attachments form.

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. Scanned areas, or 'clip image areas',

corresponded to pre-defined areas within each script in which it was expected that pupils would record their answers. These clips formed the basis of the e-marking model.

The average rate achieved per scanner, of approximately 2300 sheets per hour, was much slower than the 5760 sheets per hour that had been forecast (see Section 4.1.1). This was due to several factors. Firstly, in some cases, the size of clips was larger than originally planned to reduce the likelihood of out of clip responses. Secondly the scanners were set up to scan every clip in 256 greyscales, with special tuning for responses expected in pencil. The original intention had been to scan in greyscales only those responses where pencil was expected and scan all other responses in black and white. Regrettably, it does not appear that this divergence from the agreed procedures was picked up during the testing phase of the scanning system. This issue unfortunately also had consequences for the transmission of files to home-based markers (see Section 2.2.5). Finally the exception rate, i.e. the number of documents that were rejected because they could not be scanned using the high-speed scanner was also higher than anticipated. At approximately nine per cent this was much higher than the exception rate during the 2001 pilot. The exception rate had a significant effect on the speed of scanning in that the total script was excluded from the document count, even if only the last page of a 12 page script failed to scan. Many of the exceptions occurred because pupils were writing or doodling over the timing marks or page identification marks that had been added to the conventional test booklets.

The large number of exceptions in the 2002 pilot has serious implications for scaling up the process to deal with much larger numbers of scripts. In the future, the scanning rate could be increased by means of a greater number of machines or by investigating different scanning resolutions, settings and formats. For any future scale-up of this pilot it would be beneficial in terms of both cost and speed if the design of the test papers were to define more rigidly the space in which pupils must enter their responses. This would also need to be emphasised in accompanying administration instructions to pupils. However, even if the test booklets were redesigned and the administration instructions were altered to accommodate the implementation of new technology marking, inevitably some scanning exceptions would still occur. It is essential that procedures for dealing with exceptions are fully tested on a large scale together with associated problems such as re-collation of partly scanned scripts, etc.

The test papers were then re-packaged and despatched to the appropriate conventional marker appointed to each school for marking. Again this procedure was managed through the DWS Mail Manager program with reconciliation of script receipt and script despatch. As the scripts were still to be marked conventionally, individual pages were re-collated into complete booklets by stapling along one side with four staples (as specified by AQA).

However, this stapling was done manually and therefore not necessarily very neatly. Some conventional markers complained because, in some cases, the stapling meant that they were unable to access the space for recording their marks. Another problem encountered by conventional markers was that of cutting themselves with the staples. There were also problems because markers received the papers in a different order to that which they had expected. Due to time constraints, the papers had been sorted alphabetically within tier instead of being sorted alphabetically according to the marker list. These issues would not be relevant in a fully scaled-up system based completely on e-marking but are relevant during any further development phase.

To a large extent, delays in despatching papers to markers were due to the speed of scanning achieved. However, delays were also due to incomplete deliveries from the contractor delivering the test papers to NCS Pearson. Processing was delayed until missing packs were delivered so that all test papers from one school could be forwarded to the allocated marker at the same time. As indicated previously a turnaround of 48 hours from receipt from schools to despatch to markers was targeted. The extent to which this was achieved is reported in Section 4.1.1.1.

Once the conventional marking had taken place the papers were returned to schools. A smaller group of schools, equating to approximately 5,000 pupils' test papers were required to despatch their test papers back to NCS Pearson for the second cycle. Test papers were received from the majority of the 35 schools, more than the minimum required for evaluation purposes. During this stage the papers were again logged in and batched according to test type. The individual script barcodes from the booklet covers were then captured so that these numbers could be used to retrieve the relevant images scanned during cycle 1 from the archive. These images were then imported into either the e-PEN system or the DWS Editor system, depending on whether the clips were to be marked by expert/clerical markers or data entry markers, as described in the following section.

Following the logging in procedure the test papers were despatched to a data capture bureau (PECS), where the conventional markers' marks were captured using a double key with verify approach. Following the return of the scripts from PECS, all the materials were packaged and returned to the participating schools. Schools participating in the second cycle will receive a CD-ROM detailing item level analysis of strengths and weaknesses by individual pupils based on conventional mark information.

2.2.4 E-Marker structure

The logic of the new technology is that e-marking can operate at item level rather than script level and that markers can therefore specialise. Using the same procedure as that employed in the 2001 pilot, each test paper was sub-divided into test question items, with each item being defined as a data entry (response selection) item, a clerical item or an expert item. Data entry was used to capture those items with a limited range of pupil responses that could be compared directly with a computerised score key. Clerical items were defined as items that could be marked by non-expert markers once training in applying a mark scheme had been given. For example, these were items that required a small element of judgement but not the professional expertise of an experienced KS3 marker. Expert items were those items that required specialist subject knowledge in order to mark them. A further benefit of e-marking is that markers can focus on a smaller number of items. In the 2002 pilot, expert markers were further sub-divided into small teams, each of which was assigned to mark 20 items. The items were allocated in such a way that each team would have a range of item types and therefore approximately the same workload. For example, each team would have an item involving the use of an overlay.

All trained 2002 KS3 markers were contacted by AQA and invited to participate in the 2002 e-marking pilot as expert e-Markers. There was great interest in the pilot, with enquiries from approximately 500 markers to AQA or QCA regarding possible participation. Of these, the markers were then selected to be representative of gender and ethnicity, and their home location and availability dates were also considered. For those doing home-marking, their computer's modem and the size of its memory were taken into account. A Chief e-Marker, two Senior e-Markers, 26 e-Markers and four reserve e-Markers were appointed by NCS Pearson following the recruitment process. The 26 markers were divided into a home-based team and a centre-based team, with one Senior e-Marker to supervise each team. Thus half of the markers attended the NCS Pearson UK headquarters at Hellaby for a period of on-site marking, whereas the others marked from home via a secure Internet connection.

Although a few markers had marked KS3 conventional test papers for the first time in 2002, most of the markers appointed were highly experienced markers, team leaders and, in a few cases, senior markers. As reported in Section 3 of this report, it is a failing of the pilot in terms of representativeness that the marking team was not representative of the marking body as a whole. The ways in which these experienced markers adapted to the e-

PEN system and the e-marking process more generally may not be typical of the marking community in general.

Five clerical markers were appointed by NCS Pearson, using temporary staff from the local area. All were graduates with A-level mathematics. Again, in terms of representation, the clerical markers were more qualified than could be expected in a scaled-up operation. Initially these markers were employed to carry out clerical marking in weeks one and three and data entry marking in weeks two and four. However, both the home-based-expert marking and clerical marking took longer than anticipated. As a result, the original five clerical markers spent a greater proportion of their time carrying out the clerical marking and a further two markers were employed to focus exclusively on data entry. Supervision of the clerical markers and data entry operators was carried out by the NCS Pearson Examinations Manager. Clerical and data entry operators worked from 9am to 5pm with a half hour lunch break and two 15 minute breaks, one in the morning and one in the afternoon. During week two, a few items originally categorised as expert items were re-classified as clerical items in order to reduce the backlog of home-based expert marking. Although this was done in consultation with the Chief e-Marker, this raises the issue as to whether expediency was adequate justification for re-classification. In any scale-up, the criteria for re-classifying items and the responsibility for making such decisions should be clearly defined.

2.2.5 E-marking

The e-marking took place from 15 July to 9 August 2002. The home-based marking commenced on 15 July, following supervisor training on 12 July for the Chief e-Marker and the home-based Senior e-Marker and one day of training on 13 July for the home-based e-Markers. Centre-based supervisor training took place on 26 July, with training for the e-Markers on 29 July, followed immediately by the centre-based e-marking. All training was held at the NCS Pearson Hellaby site. Initially, both periods of marking were scheduled to be completed within two weeks. However, as the home-based marking took considerably longer than originally anticipated, there was an overlap of marking during the third week. Following the completion of the centre-based e-marking, four expert markers carried out both conventional and e-marking of a further 100 scripts as part of a controlled evaluation of marking reliability (see Section 4.3).

In 2001, the marking was carried out using Netgrade software. In the 2002 pilot, the clerical and expert markers used the e-PEN system and data entry was captured within the DWS Editor system. The requirements for the latest version of the e-PEN system that was to be used in the pilot (including the home-based element) were specified in

November 2001. Delivery of the system was promised by NCS USA to NCS Pearson (UK) for 1 May 2002. In the event the system was not available until 13 June. This delayed testing the system and solving any teething issues before the commencement of the pilot. The main elements of the system were, fortunately, fully functional. However, some required elements behaved rather differently than had been anticipated, for example the on-screen practice and qualification elements did not replicate the live system. As a result, they were not used during training as originally planned. The intention of the qualification and practice modes was to allow markers to log on to a qualification or practice set of an item that they had been allocated to mark, using a database of items scanned in from Cycle 1 but not being used for Cycle 2 of the project. Completion of the qualification and practice set would have resulted in a report showing their success or otherwise in marking the pre-marked items. Those markers meeting the defined passing criteria would then be allowed to mark live items. If this system is to be used in any future scale-up it is important to bear in mind whether both markers and supervisors consider this to be an appropriate and acceptable equivalent to the methods used currently in conventional marking. As these aspects were not used in the pilots as anticipated it was impossible to evaluate these aspect of the e-PEN system.

In addition to problems with the practice and qualification modes, the calibration mode was also found not to be working fully and could not be used. The intended use of the calibration mode was that at certain points during marking every marker could be sent a calibration item, pre-marked by the Chief e-Marker, in order to assess marking consistency to the mark scheme. Unlike 'hidden' validity items, which will be described in Section 2.2.7, the marker receives an on-screen report showing their success or otherwise in marking the calibration item. Again, it is disappointing that this element of the system could not be evaluated either by the supervisors or by the evaluation team. Also there was no opportunity to assess the markers' attitudes to this aspect of on-going monitoring. For example, reactions to calibration items may well have been different from reactions to validity items. The failure to have time to test and refine these elements of training and monitoring before the live phase and evaluate their use in the pilot is very unfortunate. It is important in future that the required aspects of functionality are fully tested and operational during the pilot. The evaluation of the calibration mode would be particularly relevant if a less than 100 per cent double marking model were to be adopted in future.

In the e-PEN system, markers are allocated items to mark according to their marker status, expert or clerical. In the 2002 KS3 mathematics pilot, markers could generally choose from a limited range of items (approximately four) at any one time. All items, irrespective of type, were double marked, i.e. each pupil response was marked by two

different markers. The adjudication of unmatched responses was performed by the Senior e-Markers, Chief e-Marker or NCS Pearson Examinations Manager. Having chosen an item to mark, that item was presented, clip by clip, until all the clips of that item were exhausted or the marker chose to switch to another item or to log out of the system.

At item level, the first clip in the database is sent to the first logged in marker, the second response to the second and so on. In the meantime, a second set of responses are cached so that each marker has another clip waiting for them. However, if some markers are marking faster than the others, they will get the next available responses into their queue before the others. Once a marker has marked a response it is moved into a second queue, the reliability or second marking queue. The only difference in the distribution of this queue is that the system knows which responses have been marked by a specific marker. The system will not allow a response in the second queue to be routed to the same marker.

Several general issues occurred in relation to all types of marking. More specific issues are dealt with in the relevant sections that follow. Generally, the main elements of the system functioned satisfactorily and most of the markers appeared to be satisfied with the e-PEN system, although there were concerns about the speed, i.e. the time taken between the submission of a score and the presentation of the next clip. All the clerical and expert markers remarked that the system periodically ran noticeably slower and occasionally froze temporarily. This seemed to depend on the type of item being marked, the number of markers on the system and whether reports were being generated or refreshed. Sometimes markers had to log out of the system and immediately log back in, disrupting their marking. Several markers in both the clerical and expert categories reported that they could have marked somewhat faster than the system allowed, even when the system was working at 'normal' pace. There were some other minor technical problems. On one occasion, all the clerical markers were presented with a blank screen when an image failed to load and the unmarked item was sent to each marker in turn, preventing them from accessing any further items in the queue. More seriously, a problem with the server on the Sunday of week two resulted in home-based markers being unable to access the system for six hours. More specific problems relating to the home-based markers are described in the relevant section below.

The most common request by both clerical and expert markers was a facility to recall the last marked item. All markers reported that they had made errors and realised too late, i.e. at the moment of submitting a score. Both markers and supervisors found this inability to change their last submitted score frustrating and detracted from their satisfaction with marking. As all items were double-marked, the supervisors picked up such errors through adjudication, except of course where two markers coincidentally made the same error.

However, a cancel or recall button would have reduced this particular supervisory workload considerably. Also, if supervisors themselves made errors when adjudicating, they could not recall items either and these errors could then only be rectified by back-reading of their adjudications by the Chief e-Marker. A recall facility would be essential if a less than 100 per cent double-marking model were to be adopted. Although it is unfortunate that this facility was not available for the 2002 pilot, NCS Pearson have said that this will be available in future versions of e-PEN. Reviewing or changing more than the last item was an issue for expert markers and is discussed in the relevant section below.

Another problem common across all types of markers was the distinction between indicating the blank (BL) mark box for a missing response and zero for an incorrect response. Making this distinction was different from established practice for expert markers, as missing responses are marked as incorrect in conventional marking. This initially caused much confusion and led to many items being sent for adjudication when one marker had indicated a blank and the other a zero. Although several markers queried the necessity for the blank option, this would be an important distinction in the e-marking of pre-tests (and the evaluation of live tests), as missing data can often yield important information about the accessibility of particular questions and whether the length of test is appropriate.

Marker performance data for individual markers were not available on-line, although the Examinations Manager compiled some data on the performance of the various teams and sent/gave these at regular intervals to the markers. In the marking centre, this information consisted of the number of responses available to the team and the current number that had been first marked and second marked. Expert markers, particularly those working at home, found it very difficult to gauge the pace at which they were working compared to conventional marking. Several markers suggested that a personal item count on screen would have been useful and would have provided some 'reward' or motivation. Centre-based markers reported that knowing how many clips of each item were outstanding for their team gave them an incentive to focus on completing particular items. However, some markers would have liked information on their individual performance, for example the accuracy of their marking. Careful consideration needs to be given to the type of information that might be provided in future and the extent to which this is personalised and confidential. Markers will have different amounts of time available and some items will take much longer to mark. Thus information provided will need to be sensitive to such differences, particularly any comparisons with other markers or teams in respect of speed or accuracy.

One problem with the system that caused considerable problems for all markers at the beginning of their marking period was the zoom function. When markers switched to a new question it automatically loaded at a high magnification, with only part of the clip visible. Both expert and clerical markers would often mark items as blank because nothing could be seen on screen, before realising that they needed to change the magnification to resize the clip. Once adjusted, the size would remain stable until the zoom function was used again or the marker moved to a new question. At the end of the first two weeks of the home-based marking, the supervisor estimated that failure to resize the clip was the major reason for discrepancies between markers. A useful improvement would be for the lowest magnification to appear as the default setting for each new item. In the 2001 study, the issue of magnification was identified as both a beneficial aspect of the software but also a potential threat, in that viewing responses at different magnifications could conceivably impact on the nature of the judgmental processes involved.

Finally, all markers expressed concerns about the number of 'beyond the clip image area' (BtCIA) responses they received. As detailed in Section 2.3 it is an issue that must be addressed in any scale-up of the current pilot.

2.2.5.1 Expert marking

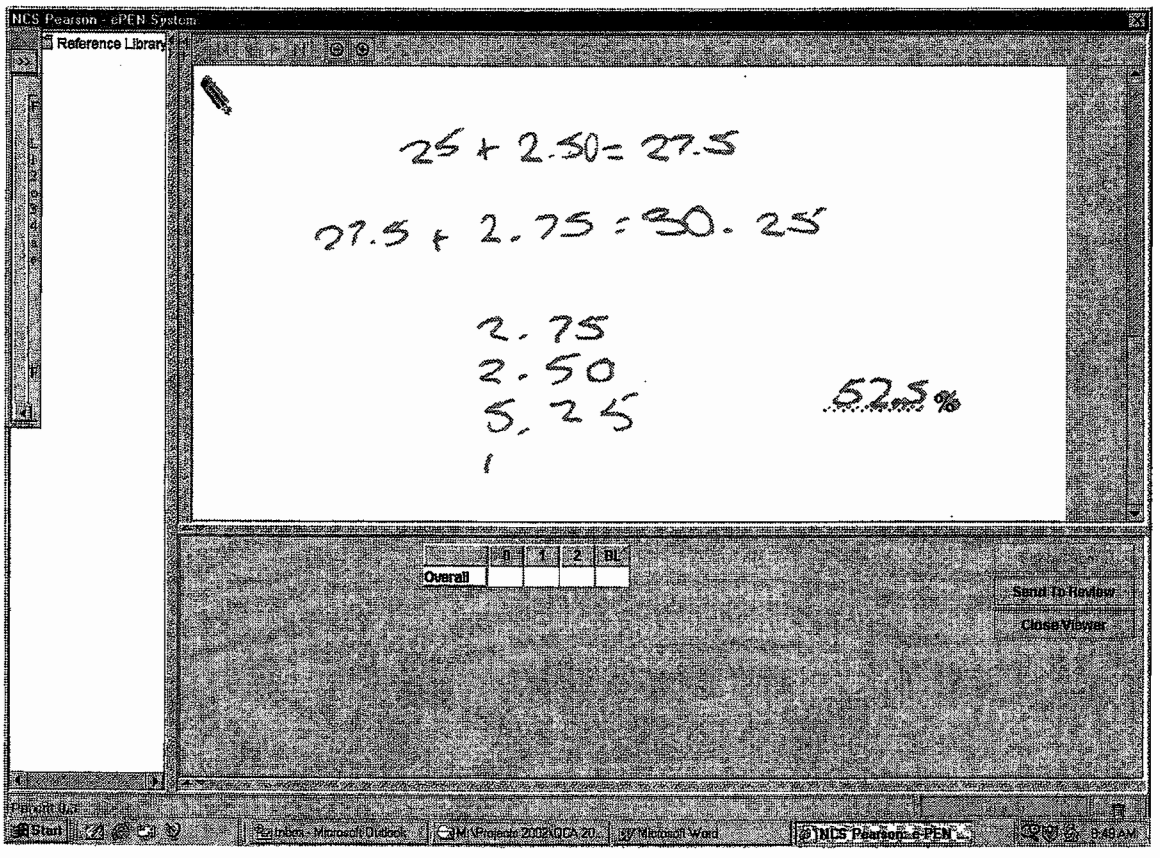
All expert markers, whether home- or centre-based, were given refresher training on the conventional mark scheme and were trained on the e-marking system on site at the marking centre.

Refresher training was focussed on just those items that expert markers had been assigned to mark within their team. In advance of the meeting markers had been sent a booklet containing approximately four pupil responses for each item that they would be marking. These items had been picked by the Chief Marker and Senior Markers to illustrate particular aspects of the mark scheme and likely pupil responses, particularly borderline or minimal responses. These marked booklets were used as the basis for the refresher training sessions and any items where there was marking disagreement were discussed.

The markers were then each given a User Guide and a demonstration of the e-PEN system on a projected screen image of the live system. As has already been mentioned, the practice, training and online qualification systems were not functioning as had been expected. The markers were therefore given to time to practice marking using a back-up database of the live items. Most of the markers seemed familiar with general computer processes such as logging on, using drop-down menus and working with windows. Once

markers had logged into the e-PEN system, selected an item and opted to 'Score student responses' they were presented with the basic screen layout as reproduced in Figure 2.1.

Figure 2.1 e-PEN screen layout for expert and clerical markers



As can be seen above, the mark boxes below the pupil response indicated the mark range available for each item. Markers were unable to submit a score unless one of the mark boxes had been selected. This was an improvement on the Netgrade system used in the 2001 pilot, where markers could commit a score without selecting a mark, resulting in a no-response default mark. Where items needed to be marked together, due to the follow through nature of the mark scheme, the mark boxes were shown together on the same screen and all had to be completed before the item scores could be submitted. Where markers were unsure how to mark items, they could use the 'send to review' button to forward them to a supervisor. In such cases, they could either award a mark or leave the mark boxes blank. Using the 'send to review' button gave them the option of specifying whether the query was a mark scheme issue, suspected malpractice, out of clip (BtCIA) or other (usually illegible responses) and to add a message to accompany the item if they wished to do so. No data were made available quantifying the reasons why items were sent for review. The mark scheme for the item currently being marked was available in

the left-hand window and could be opened and closed as required. The pupil response window and the mark scheme window could be resized as required.

2.2.5.1.1 Home based markers

Initially, the home-based markers were contracted to mark on-line for 40 hours over a two-week period. In the first week, the group comprised 13 markers and two reserve markers. However, because it soon became apparent that the home-based marking would take longer than originally anticipated, the two reserve markers and a further six reserve markers were also contracted to work at home. The home-based database was kept open for a third week for those markers able to continue marking in an attempt to complete all the items. The assumption that markers would mark 275 items per hour was an over estimate and the actual average was only approximately half that rate. In any scale-up of this pilot, consideration must be given to the speed of the home-based marking achieved in 2002 and the implications of this in terms of both recruiting sufficient home-based markers and meeting the same deadlines as conventional marking.

Home-based markers were divided into four main teams, with one extra team being created in the second week (this team was made up of the reserve markers). Markers were allocated into teams somewhat randomly, but with an attempt to create a mix of experience. It was originally thought that each team would mark 20 different items. However, the main home-based teams ended up marking an average of 17 different items, with the extra team marking eight items.

All of the home-based markers installed the e-PEN software onto their home PCs using the 'Home Based Markers Installation CD' and accompanying instructions (version 3.4). All home-based markers fulfilled the minimum software requirement, which were set at a modem running at 56K, and a PC with 128MB of memory.

Generally the markers found the e-PEN software easy to install. However, installation times of between 20 minutes to 2 hours were reported. For those markers who reported times at the upper end of this range, this had been mainly due to up grading their PC with appropriate software upgrades included on the installation CD. A few markers experienced slight technical problems entering the e-PEN website address, as the instructions gave 'https' when users should have entered 'http'. However this problem was quickly resolved by the technical helpline and all the markers successfully installed the e-PEN software in time to start the live marking.

Although the home-based marker training day was not directly observed, it is presumed that it followed a similar format to that of the centre-based training day. Many of the

markers who were interviewed made positive comments about the training day. The supervisor reported that face to face training was essential for home-based markers, followed by on-going telephone support.

The majority of markers reported that the e-PEN system was easy to use and navigate through, although some markers made specific criticisms in reference to the screen layout. Most frequently markers commented upon the positioning of the mark and the submit score buttons, the general feeling being that the two buttons were too far apart. One marker even thought the distance between the two buttons could cause or aggravate a repetitive strain injury. Despite these general criticisms, one marker commented that the distance between the two buttons was useful as it gave markers time to think, before they finally submitted their mark. Of those who criticised the screen layout, many thought it would be more convenient to reposition the mark and submit score buttons vertically at the right of the screen. A few other markers thought that the screen layout should be adapted so that pupil responses were given more room. Other than these specific issues relating to the screen layout, many markers said that they had found the screen layout reasonably flexible and sufficient to carry out on-line marking.

Criticisms were also made in reference to the on-line mark scheme. Although nearly all of the home-based markers had used the on-line mark scheme, many continued to rely upon their paper copy. Of those markers who had used the on-line mark scheme several felt that the annotation function was time consuming and inadequate. Many markers thought annotations should have been automatically visible when the rest of the mark scheme was being viewed. Similarly markers wanted to be able to highlight and colour code the on-line mark scheme in the same way that they would have personalised their paper copy. A general concern was that if markers are unable to highlight their mark schemes in this way, then some of the finer points of the mark scheme may be lost amongst the sheer volume of information being presented on screen. Several of the markers were concerned that less senior markers may not successfully distinguish between examples of correct and incorrect answers, which they reported as a frequent cause of error in the paper marking. In addition to these comments, markers also criticised the length of time it had taken to scroll through the on-line mark scheme, to find relevant information. Generally, markers thought that the on-line mark scheme would have been more useful if they had been able to view a whole page at a time. However, some markers had found the on-line mark scheme particularly useful for marking short responses. They reported being able to mark short response items more quickly because they were able to position the mark scheme and the response alongside each other. Many

markers also suggested that on-line marking would have been easier if the question they were marking either accompanied the response or could have been located on the screen.

Many of the markers said that, having completed so much paper marking, they had already memorised a large percentage of the mark scheme and therefore did not feel they needed to refer to either paper or on-screen mark scheme very frequently.

As has already been mentioned, markers were also frustrated by the time taken between a marked response being submitted and a new response being presented. Some markers reported maximum submission times of up to 60 seconds. The delays in submission times were apparently caused by the use of greyscale and a high scanning resolution (240x240dpi), resulting in large file sizes of images (in one case up to 800KB). This problem was detected part way through the marking period. Image resolutions were reduced to 120x120dpi on any files of 100KB or more to reduce image files to a third of their original size. This measure was partly successful and some markers did report that submission times had improved slightly during the second week of marking. However, generally markers were somewhat frustrated by this problem and felt that lengthy submission times were slowing down their overall performance. Many thought this would have been an even greater concern if they had been paid per marked item. This problem could have been prevented had the system been sufficiently tested prior to the live pilot.

Some of the home-based markers also felt that the submission times were being affected by global internet access. For example, the system was reported as particularly slow during the mid morning when the internet is at its busiest. One marker in particular found that during these times a 'timed out' message appeared on their PC and access to the e-PEN site was denied. The help line was unable to resolve this problem and they were only able to suggest that the marker worked during less busy popular times of day. Consequently the marker had to start marking before 7.00am to guarantee access to the e-PEN system. However, some markers did report faster submission times. For those few markers who had a broadband internet connection and were not relying upon a 56k cable modem, substantially faster submission times were reported. This study reveals that, although it was possible to mark using a 56k modem, this was very slow and therefore an inefficient use of expert markers' time. Certainly many markers felt that if they were to continue to take part in on-line marking in the future then a broadband internet connection would be essential.

Due to the frustrating submission delays, several markers suggested that an audible beep would be a useful adaptation to alert markers when a new response was available to mark.

Other markers thought that submission times might be reduced if responses could be sent in batches to their PC.

In addition to problems with the speed of item delivery, nearly all markers experienced the system freezing. In some cases the system corrected itself and in others markers used a combination of Ctrl/Alt/Delete or restart to unfreeze the system. Only in a few cases was continual freezing an unresolved problem. One marker in particular found that the system froze between six to eight times per hour. This marker did not experience a continued period of marking that extended beyond 15 minutes. The technical support helpline thought that the problem might have been caused by a poor internet connection, exacerbated by living in a rural location. It is regrettable that no testing of the capabilities of the home-based markers' equipment was carried out prior to the live marking phase. In the future, provisions should be made to check that all on-line markers have sufficiently robust computers and internet connections so that similar problems do not reoccur.

However, on the whole, few markers experienced technical problems which could not be resolved. Of those markers who had sought technical support the majority were very complimentary of the service provided. Markers were particularly impressed with the availability of the staff and their punctuality in returning calls. Technical support was provided between 8am and 11pm and most markers found that these times fitted in with their marking schedule.

As with all marking, the legibility of student responses can sometimes be problematic. Generally, the home-based markers thought that student responses were as clear scanned on screen as they were on paper. In some cases, markers praised the on-line zoom function as a useful facility for clarifying a pupil's response. Only one marker felt that the legibility of pupil responses had been compromised on-line and thought that the reproduction was too faint. Similarly, one marker felt that the legibility of response was sometimes compromised on diagrams, graphs, etc. because both text and students' responses were monochrome, making it sometimes difficult to distinguish pupil responses. A few other markers felt that one disadvantage of e-marking was that they were unable to cross reference illegible or ambiguous responses with other sections of a student's script. One marker felt that marking would be faster and more accurate if all blank responses could be filtered out prior to sending responses to markers.

Many markers reported that the most frequent cause of an item being illegible was because part of the pupil response lay beyond the clip image area. As will be mentioned later in this report, BtCIA responses were the most frequently reported queries sent to the supervisor to review. Many markers made suggestions which they thought might help

reduce the high number of such responses. The majority thought that tests should be designed so that pupils were given boxes in which to write their answers. One marker thought that the scanned area could be slightly enlarged to help reduce the number of out of clip responses.

Markers gave a mixed response when asked how they felt about marking a series of the same question type rather than individual booklets. A few markers described marking a limited number of question types as 'tedious'. However, many markers actually reported it to be a positive feature of on-line marking. They felt that marking a larger volume of the same question type meant that they became well acquainted with the mark scheme and some suggested that this in turn may improve marking accuracy. Conversely, one marker suggested that it may have the opposite effect, causing markers to be over confident and more error prone. The majority of markers spent between half an hour and one hour marking the same question before switching to a new question type, with the exception of one marker who only marked the same question for between 5-30 minutes. Nearly all markers reported switching between different question types as a strategy to relieve boredom. Most markers did not mark for longer than two hours before having a break and some markers took breaks more frequently than this. Markers were never without responses to mark, other than when the server went down. One marker noted that topic areas that had been completed continued to appear under available tasks. This was most probably due to the supervisor not removing the task from the menu of those available once it had been completed. In practice, a marker attempting to do this task once it had been completed would reach a display indicating that there were no remaining items.

Many markers gave specific examples of questions that had caused them particular problems. Although very few markers made reference to the same questions, many mentioned questions involving an overlay as problematic. These were the equivalent of the acetates used by conventional markers to assess the accuracy of angles drawn by pupils, and involved a computer generated graphic rendition of the range of acceptable responses which could be transposed over the clip image. Some markers had difficulties rotating the overlay, while others found the overlay difficult to size and position. It was certainly noted that, towards the end of the second week, many of the home-based markers had not attempted to mark a large proportion of the questions that required an overlay to be used. The implication for scaling up the pilot is that, unless markers are allocated a specific number of each item, they may choose not to mark problematic items, such as those involving overlays.

Another general comment made by markers was in reference to questions that included several parts or went over two pages. Sometimes markers needed to cross reference the

pupil's response on one page with a graph or a diagram upon another. Some markers reported that diagrams and graphs, which may have included extra information, had not always been included in the clipped area. A similar criticism was made about questions that were divided into two parts and had been clipped separately. Some markers found that they were unable to mark the second part of the question because it made reference to information contained within the first. This indicates that decisions about the size of clips and to what extent items are scanned as conjoined clips (i.e. two or more items scanned together) are extremely important to ensure that pupils do not lose marks compared to conventional marking.

Home based markers had mixed views on how they responded to awkward or difficult items. A few markers said that they deliberated for longer over awkward items, knowing that they were being double marked. Many of the markers taking part in the home-based marking were senior markers or team leaders and, because of this, they reported that they had felt it was their professional duty to deliberate over items before sending them to review, as they were used to making supervisory decisions. It may be for this reason that the majority of markers reported mainly sending 'out of clip' responses to review. However, some markers did report using the 'send to review' function for submitting awkward items and several admitted using it as a way of gaining encouragement and feedback from their supervisor. Although the majority of markers said that they preferred to deliberate over items, others said that they were readily using 'send to review' as a means of avoiding deliberation.

A few markers said that they sent fewer items to review in the second week. Although this was partly due to feeling more at ease trusting their own judgement, one marker said that s/he had stopped sending items to review because there was rarely any response from the supervisor. Note, however, that the sending of replies from supervisors to every 'send to review' query was not an expected part of the review procedure. As has already been mentioned, a common concern raised by many markers was that, when supervisors did reply to a 'send to review' query, the message was sent back without the image attached. Even when supervisors were explicit in their feedback, some messages were often meaningless without the relevant clip.

In addition to the 'send to review' facility, markers were also able to send and receive feedback using e-mail and phone. All the home-based markers received some feedback from their supervisor, such as the general group e-mails, but not all markers received individual feedback. There appeared to be some disparity between the amounts of feedback given to markers by their supervisor. A few markers felt that there had been a delay in communication between supervisor and marker. Also many of the markers

would have liked more feedback detailing their individual and team performance. However, some markers thought that the level of feedback they had received had been appropriate. Very few reported seeking specific feedback to a query they had with the mark scheme and said that this had been because the mark scheme was so comprehensive. Some markers commented favourably upon the feature of the system which meant that markers had to read their messages before they could proceed with marking.

As might be expected, many of the home-based markers noted that they found the experience of on-line marking different from that of conventional marking. A number of markers reported that they found the e-PEN system frustrating to use. The greatest causes of frustration reported by home-based markers were the time taken to submit responses and being aware of making errors but not being able to recall items. As has already been mentioned, home-based markers also found that not knowing how they were performing or what progression they had made was de-motivating. Some markers said that they would have liked an on-screen device which allowed them to see how many questions they had marked. Some markers also felt that their attitude towards on-line marking was different from their attitude to conventional marking because it required a higher level of concentration.

Very few of the home-based markers had experienced problems marking within their home environment. Many reported making minor adjustments to the position of their chair, desk and PCs to maximise their comfort. Only one marker reported severe discomfort and this was during his first week of marking, when he relied upon his lap top finger mouse. It may need to be specified in the future that this type of mouse is unsuitable for on-line marking. Generally, markers felt that computer marking was less physically demanding than conventional marking, as they felt they did not have to move paper around or be hunched over scripts. One marker was also pleased to note that on-line marking did not discriminate against left handed markers. Few markers reported that completing on-line marking had caused any major disruption to themselves or any other member of their household. The most frequently made comment was with regard to the installation of a second phone line. Many markers felt that if they were to conduct future home-based on-line marking then the installation of a second phone line would be essential.

Comparing e-marking with conventional marking, some home-based markers reported slight differences in their working patterns. A few markers thought that they had not spent as long on the on-line marking as they would have done on conventional paper marking. Conversely, some markers said that they marked for longer on-line because they enjoyed the computer based marking, although they thought this might just be because it

was a novel procedure. The majority of markers reported taking regular short breaks, though generally they did not think they were taking them any more frequently than when completing conventional marking.

2.2.5.1.2 Centre based markers

As the home-based marking had taken longer than anticipated, it had been decided to use 17 markers in the marking centre. This included the original team of 13 markers plus three reserve markers and a marker who had been assigned to the home-based team, but in the event had been unable to carry out any marking at home. Markers were encouraged to mark more than their contracted 42 hours if they wished to do so, in an attempt to ensure that the marking was completed within the specified time period. Markers were able to mark any time between 8am and 8pm and some agreed to mark on the Saturday of the first week. The number of hours worked per day varied from marker to marker, depending to some extent on whether they were staying in Hellaby or travelling home each evening. During the second week, three of the home-based markers came into the marker centre to replace centre-based markers who were unable to work that week. The aim was to try and keep 15 workstations operational each day. Prior to the marking period, markers were allocated into one of four teams, each of which had been assigned 20 items to mark. (Please refer to page 24 for a definition of 'items'.)

At the beginning of the introductory session on day one, markers were given a brief overview of the project, general administrative instructions, and a leaflet on health and safety. The overview included an explanation of the three types of item: data entry, clerical and expert, the use of validity items and the use of double marking. Markers had previously been allocated to one of four teams and each team was to be given approximately four questions at a time to mark. The Examinations Manager had allocated the different types of questions across the teams in order to try and give similar workloads to each team. As outlined previously, markers received refresher training on the items they were going to mark, training on the e-PEN system and a period of e-PEN practice. Production marking then commenced at approximately 3pm on day one of the centre-based marking period. Markers could chose to work at any of the designated workstations, linked to the NCS Pearson UK intranet, and take formal or informal breaks as and when they wished to do so.

During refresher training in one group, a couple of general issues were discussed. One was that a marker in the home-based group had realised that errors had been made during conventional marking because s/he had 'missed' something on the mark scheme. It was suggested by the trainer that a possible benefit of marking a smaller number of items was

that markers would be able to get to know a smaller section of the mark scheme extremely well. It was also felt that markers would see a wider range of responses because items would not come from a small number of schools. Similar comments were made later in the week by markers during interviews with the evaluation team. As with the home-based markers, the majority were in favour of marking a smaller number of items as they commented that they would be both quicker and more consistent when they were focusing on fewer items.

During e-PEN training, there was an explanation of some of the issues that had caused problems during the early stages of the home-based marking phase. These were: the need to differentiate between missing and incorrect; the need to reduce the magnification of the screen on new items; and the need to enter the question number in the messaging box if feedback is required when sending items for review. However, this last item could not be fully demonstrated on the projected screen image of the system, as there appeared to be a minor fault with the system at that point in the demonstration. The various sections of the mark scheme were introduced and markers were requested to try and use this for at least part of the time so that they could feedback their views on this part of the system.

As has already been reported, markers at the marking centre were generally satisfied with the e-PEN system, found the software easy to use and had generally enjoyed the experience, even those who initially had reservations. The following section highlights the most frequent comments that emerged during event shadowing and individual interviews.

Generally markers liked the Windows environment and adjusted the layout to their individual preferences. Some remarked that they liked the facility to confirm their score with the submit score button. However, a few markers commented that the mark boxes were too close together and would occasionally cause them to hit the wrong box. They asked for the marking boxes to be separated on the screen and also for the blank response (BL) to be situated next to the incorrect response (0). The majority of markers asked for the questions themselves to be shown somewhere on screen – this was reported as being especially useful for certain questions, such as algebra. Discussion with NCS Pearson indicated that questions were not included on screen to save space and file size. In future, it is important that the test papers are pilot marked to identify any questions where having the question visible on screen is likely to significantly increase marking accuracy. Most markers were pleased with the screen resolution and several remarked that the facility to magnify the clips made it much easier to see pupils' responses than in conventional marking. The images for centre-based markers used the same 120x120dpi, 256-greyscale resolution to which the home-based images had been converted. The home-based

supervisor commented that a lighten/darken icon on his display was useful for detecting crossing out but noted that markers did not have this functionality.

All of the markers had tried to use the on-line mark scheme for at least part of the time and centre-based markers made similar comments to those of the home-based expert markers. For example, the reported usefulness of the on-line mark scheme tended to depend on the type of items being marked as well as individual preferences. For items with diagrams or less detailed marking criteria, several markers reported that they had used the on-line mark scheme, positioning a suitably sized window alongside the pupil responses. However, where the mark scheme was particularly detailed, some markers found it time consuming scrolling up and down to find the appropriate part of the mark scheme and reported that it was easier to glance at a whole page of hard copy. A few markers commented that it would be useful to be able to check the on-line mark scheme for updates before beginning to mark a new item. Several markers wanted the facility to highlight or annotate the mark scheme and one commented that markers would only feel comfortable with the on-line version if they were able to make notes during training and 'take ownership' of it. One marker had tried to add some additional notes to the on-line mark scheme but had found that some mathematical notation was not recognised. Another marker requested an on screen calculator. A common complaint was that it was not possible to have both the overlay and mark scheme open at same time. One marker commented that often the overlay questions involved deciding on a checklist of several marks for different features. It was essential to be able to glance at both the mark scheme and the pupil's response with the overlay applied. Another problem with the overlays was that when one score was submitted the overlay did not refresh immediately. Often the overlay became opaque or retained the last image. Markers sometimes marked the same item twice not realising that the overlay had not refreshed. Each time a new item appeared markers had to click on the image or on the refresh icon, before being able to mark the next item, thus wasting time. Adjusting the overlays was also time consuming although markers did not seem to mind spending a few minutes doing this if they were going to be marking this item for some time. A few markers suggested that it would be better if the overlay could be anchored to the scanned image and then stretched to resize if necessary.

As has already been mentioned, all markers wanted to be able to recall the last submitted score to check or amend their last mark. Some expert markers would also have liked the facility to go back several items. They reported that, if they recognise they have made an error in conventional marking, they will often go back through several previous scripts to check they have not made a similar error previously. Alternatively, a team leader may ask them to go back and remark some scripts following clarification of the mark scheme. In

the 2001 pilot the Netgrade software allowed markers to return to any of the previous ten items marked and amend the marks they had awarded using the 'back score' function. A similar functionality within e-PEN would be useful and necessary, particularly if scripts were not 100 per cent double-marked. If it were not possible to recall more than one item in a scaled-up operation, on-line re-marking would have to be dealt with in some other way than in conventional marking.

Working conditions were generally considered to be good, with spacious well-lit areas, drinking water available and an eating area with snack food facilities. Although there were some individual comments, most markers were happy with the workstations, equipment provided and working environment. Some mentioned that the two horseshoe-shaped workstations for three people were slightly cramped for comfortable working and preferred the more spacious workstations for four people. Two markers thought that wrist supports would be beneficial due to the amount of mouse use involved and one that presenting the mark boxes vertically on the right hand side of screen would reduce rotation of the wrist. Several markers described how they had found it necessary to take frequent very short breaks away from the screen, especially if they were marking questions with overlays requiring more focusing on the screen. One marker was concerned that the need for such breaks should be acknowledged within whatever type of pay scheme is developed if this pilot is scaled-up further. Although markers were given a health and safety leaflet on working with VDUs, no information had been requested by NCS Pearson in advance of the marking period about any particular workstation requirements and no workplace assessments were carried out. It is recommended that more consideration of health and safety issues should be taken into account in any scale-up of the pilot.

Although one or two markers had found it somewhat boring, the vast majority of markers were happy to mark fewer questions. Several reported that, despite their initial reservations, they had found it less tedious than expected. A few had found it more interesting because they had seen a much wider range of responses on some of the items they had marked than they would have done during conventional marking. The anticipated advantages over conventional marking most commonly reported were getting to know part of the mark scheme very well, requiring less training and promoting more accuracy and consistency. One marker commented that e-marking judgements would not be affected by previous responses on the paper or by the tier of entry (i.e. the 'halo effect'). Several markers also commented that one of the major benefits of e-marking was that it would be a more efficient use of their time and would relieve them of clerical/administrative tasks, such as totalling scores, and the checking, packing and despatch of scripts. One marker felt that it would be much easier and quicker to send

items for review using an on-line system than the current system, whereby markers have to telephone or e-mail for advice, and that this would also increase accuracy and consistency between markers.

Markers adopted different strategies in respect of how long they marked one particular question. Some were happy to mark the same question all day whereas others marked one question for no more than an hour and then swapped to a different question. The majority appeared to work on a question for two to three hours at a time, although this also tended to depend on the particular item being marked.

When asked about their opinions of the messaging system, many markers reported that they had sent items for review but had not received any feedback electronically. When questioned further, it appeared that many of these items were BtCIA responses. Other common instances were where they had marked an item but felt it was necessary to explain why they had awarded or not awarded a mark. Although most markers had assumed that they would only receive a reply if the supervisor had disagreed with their decision, there was no way for them to know whether the supervisor had looked at these review items or not. One marker suggested that, if he were working at home, he would still tend to telephone for clarification on marking points if he needed it urgently. If the electronic system is to be used for this, there needs to be some means for markers to highlight those items where they need urgent feedback, for example clarification of a marking issue, so that supervisors can prioritise these. In the marking centre the supervisor did not use the review messaging system because she preferred to give personal feedback. She would either go over to the marker concerned to talk to them or ask them to come and discuss an item on her screen that had been sent for adjudication or review. Most markers commented that personal supervision was preferable to on-line messages as it was more immediate. Regarding supervision generally, two markers felt that there should have been more supervisors for the number of markers at the marking centre. Additionally, one of these markers commented that markers would be very unhappy if they suspected that one intention of a move to e-marking were a reduction in the level of supervision.

Although the majority of markers' comments tended to focus on the advantages of e-marking over conventional marking, two specific disadvantages were mentioned by a few markers. Firstly, as reported by the home-based markers, it was not possible to look at the handwriting or the pupil's working across the whole script, to assist markers in making decisions about potentially ambiguous responses. Also, it would be almost impossible to pick up instances of copying or cheating that can be spotted in conventional marking when markers see identical answers over a range of questions in consecutive scripts. A

couple of markers expressed concerns about issues of follow-through and felt that complete questions should be kept together and not separated into separate clips for marking.

With regard to the possibility of e-marking in the future, preferences for working at home as opposed to working in a marking centre were fairly evenly split. However, interestingly, some markers who initially expressed a preference for working at home, altered their opinion over the course of the marking period. They had enjoyed working in teams and the fact that they could discuss marking issues and resolve them with their colleagues in a supportive atmosphere. This could be observed in that team members increasingly chose to sit together when marking and discussed progress on items and 'tactics' for those items yet to be marked. Two mentioned that there were fewer distractions in the centre and they could stay on task more easily. Other markers could see the advantages of a centre but would have to work at home if the marking were earlier in the year because of other work commitments. Several markers reported that their ability to mark in a centre would depend on the convenience of the location and the flexibility of the opening hours. Some markers would definitely prefer to work from home in future but had doubts about the capabilities of their home computer system and/or the speed of their internet connection. Any scale-up of the home-based pilot would have to address the issue of who would be responsible for funding or providing the necessary equipment and technical support for markers wishing to mark from home. Those markers who had marked both at home and in the centre all reported faster submission/download speeds when working in the marking centre. They also commented upon the advantages of having fewer distractions and being able to discuss marking issues with colleagues and receive immediate face-to-face supervision when required.

During their interviews, several markers asked questions about the implications or potential arrangements for paying markers if e-marking were to be scaled-up in the future. Although there was great interest in marking during the pilot, such questions revealed that, in the long-term, any system of payment must be perceived as equitable if sufficient markers are to be recruited. Some markers reported that they felt obliged to work harder because they were being paid by the hour but noted that not all markers appeared to be working at the same rate. This could be a potential source of friction in a marking centre. However, if a decision were made to pay by the item this would raise other concerns. As some questions are inevitably going to be more difficult and therefore slower to mark than others, there were obviously concerns as to how questions would be allocated, the extent to which markers would be allowed to choose which items to mark and how payment

might reflect such factors. Another view expressed was that the e-marking system had the potential for implementing a payment system that rewarded accuracy.

Overall, the centre-based markers were favourably disposed to e-marking and were very positive about taking the idea forward, particularly if some of the technical issues, such as the speed of downloading and the overlay technology, can be improved. Several asked whether they would be informed about the outcomes of this pilot and commented that they would like to be involved in any future development work if possible.

2.2.5.2 Clerical marking

The clerical markers required training in both the mark scheme and the marking system. Training on the mark scheme involved a discussion of general guidance issues, supported by a single-sided A4 sheet of notes, e.g. how to deal with transcription errors. Some instructions, such as the distinction between missing and incorrect and the need to resize the screen for new items, were given verbally but were not included in the written instructions. Markers were then given an e-PEN User Guide and were to be given a demonstration of the e-PEN system in the conference room using a projected computer image. However, due to technical problems this was not possible and the training was based on an explanation of the screens reproduced in the User Guide. After this initial session, the markers were taken through to the marking area. For these markers a demonstration was given on the live system, with the result that these markers had no experience of the system until working on live items. The screen layout and functionality were the same as those illustrated earlier for expert markers.

Before marking on screen, the clerical markers were given paper-based training on individual items. This involved an explanation of the marking criteria, followed by each marker working individually through the same six demonstration items, deciding how they would mark them. These demonstration items had been devised by the Examinations Manager, in conjunction with the Chief e-Marker, to exemplify the full range of acceptable responses. Once these answers had been discussed within the group, the markers went to their workstations and began to mark live items on screen. In discussions during training it was emphasised that, as all items were double-marked, errors would be picked up by discrepancies between markers. Initially clerical markers were trained on one question at a time. Later they were trained on several questions at once so that they could move straight from one question to the next without waiting for more training each time. Clerical markers marked all the clips for one item until there were no items to mark. They then moved to the next item on their list and repeated this process.

Although generally the training appeared to be straightforward and understood by the clerical markers, it was observed that there were occasionally some limitations to the paper-based training materials. The training materials did not include the questions themselves and in some instances the information given was not always fully comprehensive. For example, in a question where pupils had to show a specific direction on a grid, the training focused on the ways in which the correct direction could be indicated. The scale of the grid was not discussed, although it may have been implied by some of the practice examples. When several of these items were later picked up as having been wrongly marked, it transpired that some markers had not understood that both the direction and the scale had to be correct. If the training is inadequate there is a possibility that several markers will make the same mistake and that these errors will not necessarily be picked up by adjudication.

Generally, the clerical markers found the system easy to use. There were occasional problems with screens freezing but most of these were easily overcome by markers logging off and then back on again. The clerical markers were advised to click regularly on a dustbin icon to clear 'dead' images, but reported that this did not noticeably increase the speed of the system. Most markers found the speed of downloading acceptable although some felt they could have marked faster than the system allowed. This became more noticeable during the expert centre-based marking when all of the clerical markers commented that the system was running much slower. Several markers commented that they would have preferred the mark boxes to be slightly larger and nearer to the submit score box to reduce the amount of mouse movement required. The most common request from all the clerical markers was to have a cancel button to recall the last score submitted. Although they were aware that double-marking would pick up errors, they reported that they would prefer to self-correct errors that they had recognised themselves.

Most of the items sent to review by the clerical markers were 'out of clip' or illegible responses. However, when items were sent to review due to marking issues, they reported that feedback was rarely useful, as by the time any message was received they had already moved on to another item. A slight concern emerged towards the very end of the marking period, when the clerical markers had got to know one another better. There was occasionally a tendency for them to discuss amongst themselves how to mark responses that did not exactly match the mark scheme rather than send items to review or ask the advice of the Examinations Manager. It is possible that such decisions may have resulted in two markers making the same marking error and that those items would not then have appeared for adjudication.

On day one of the clerical marking productivity was high, with over 700 items per hour being marked, compared to the anticipated rate of 480 items per hour. However, this rate slowed considerably as the level of difficulty of the marking increased. Originally it was anticipated that the clerical marking of the home-based items would be completed in five days, but in the event the marking took seven days. A further day was allocated to the marking of three additional items, originally categorised as expert items, which were felt to be within their marking capability.

2.2.5.3 Data entry marking

Data entry marking was located in the marking centre only. It was carried out using the DWS Editor application by the five clerical markers, plus two further markers employed specifically for this task. All of the data was input twice, with adjudication by the NCS Pearson Examinations Manager.

The DWS Editor system differed from the e-PEN system but the software appeared to function well and was user-friendly. Data entry markers entered all the responses to a small set of data entry items from pupil 1, then the same items from pupil 2, pupil 3 and so on. Images of scanned pages were displayed on screen, with the area around each pupil response highlighted in yellow. The system moved automatically from one highlighted area to the next and the markers entered exactly what was seen within that area only. Training consisted of a single A4 sheet of instructions plus a brief explanation of the system. However, the data entry system was considered by the markers to be very straightforward, easy to understand and use. Blank responses were entered with the backslash character. If the data entry markers could see a response or part response outside the yellow area, words instead of numbers, or the response was ambiguous/illegible, those responses were sent to a supervisor for review with an accompanying message. Problems with the system appeared to be relatively few and were quickly rectified, e.g. initially the backslash character was not accepted for some items. Due to a whole page being displayed together on this system, markers could scroll back to previous responses on that page and self-correct any errors that they realised they had just made.

Captured pupil responses were compared against a set of acceptable values, as defined by the Chief e-Marker, and marks were awarded automatically. This comparison and data processing was done by machine. Data entry marking was double entered to assure quality, and discrepancies sent for adjudication by a supervisor.

2.2.6 E-marker supervision

2.2.6.1 E-PEN user (clerical and expert marker) supervision

Supervision of the expert markers was carried out by the two Senior e-Markers and the Chief e-Marker, whereas supervision of the clerical e-Markers was carried out by the Examinations Manager. In the pilot, the Examinations Manager was above the Chief Marker in the hierarchy. However in a scaled-up pilot there was some uncertainty about whether the responsibilities of the Examinations Manager should be administrative only and separate from marking issues and supervisory functions. For example the Chief Marker had concerns about some of the marking decisions made by the Examinations Manager in respect of some of the pre-loaded validity items.

Only one day was allocated to supervisor training and this was carried out only one day in advance of the marker training. There was no 'dummy' data to practice using the system on and to make the reports for supervising and monitoring markers more meaningful. In future it is suggested that there should be considerably more time allocated to the training of supervisors, not only in the use of the system itself, but also in the procedures for supervising and monitoring their teams and their specific roles and responsibilities. As reported in the 2001 report, the supervisor's role may differ according to the marking model adopted and the marking hierarchy and therefore training should be tailored accordingly. The exact role to be adopted in the 2002 pilot was not clearly defined, resulting in confusion as to whether supervisors should be primarily amending incorrect answers (quality control) or providing formative feedback to markers (quality assurance).

As a result of the experience of the home-based supervisor, the Centre-Based Supervisor User Guide had been updated to include some 'how to' sections for the main supervisory functions. The training involved a demonstration of the expert marking process followed by a demonstration of the supervisory functions. As the home-based supervisor and Chief e-Marker were also present, it also served as refresher training for them. Gaps in their understanding indicated that some of their initial training prior to the home-based marking period had been insufficient or ineffective.

One of the major problems experienced by the home-based supervisor was identifying the number of responses requiring adjudication or awaiting review for each item. This information would then allow him to prioritise those items that were producing the greatest number of queries or unmatched responses. Unfortunately, these vital pieces of information were not available in a accessible or user-friendly format within the e-PEN reporting menu. For example, the completion report listed items by their clip name whereas the adjudication screen listed items by their question code. On the advice of

NCS Pearson, the home-based supervisor eventually used the *Queue Status* report, from another e-PEN website, to obtain this information. Based on the experience of the home-based supervisor during the first two weeks of marking, this particular problem was recognised by NCS Pearson. Separate reports were produced regularly (approximately every 2-4 hours) for the centre-based supervisor, listing the name, item number and quantity of items awaiting adjudication and review. These could not be viewed by the supervisor on the e-PEN system but were provided in hard copy. As the reports were presented in a clear bar chart format, the supervisor could very quickly see at a glance where there were large numbers of items requiring attention. In any future pilots, it would be advisable to use one common descriptor across all the marking and supervisory systems. However such descriptors would need to be linked to a simple numeric identifier for data analysis purposes.

In the pilot, the relatively flat structure of the marking team did not replicate the full hierarchy of the conventional teams approach. In a scaled-up operation it would be important to test exactly how such a hierarchy would work in practice. Both of the supervisors in the pilot felt that they had too many markers to supervise effectively and that smaller teams would need to be created in future. The home-based supervisor reported that even the initial ratio of 13:1 gave him insufficient time to deal with all the reviews and adjudication in sufficient time make a difference on the rest of the marking on that item. Considering that the majority of the markers in this pilot were experienced and relatively senior members within the conventional marking system, the true workload of supervisors in a scaled-up pilot may be even higher. In the 2002 pilot, the intention was that the Chief e-Marker would be able to monitor the decisions of markers and supervisors by means of 'back–reading'. Any changes made within the back-reading mode had the effect of overriding all previous marks including first mark, second mark and adjudicator marks. However, in practice, particularly in the first two weeks of marking, the Chief e-Marker found he had insufficient time for carrying out this monitoring role as he was spending much of his time assisting with the adjudication of home-based markers.

There was some concern that there should be sufficient time to check the quality of marking before markers proceeded to mark hundreds or thousands of clips. In particular, with home-based marking, where markers could mark at any time of day or night, there could be considerable time delays between a marker starting to mark and his/her supervisor examining information about the quality of that marking. It was suggested that markers should be allowed to mark only a small number of each item until supervisors had had time to check quality of their marking (in effect replicating the system of sample scripts in conventional marking). Decisions would have to be made as to the best means

of accomplishing this, for example through back reading, adjudication or by reviewing data on reliability and validity. It is possible that this checking could have been accomplished within the training and practice modules, but, as these were not available during the 2002 pilot, it is not at all clear how these would have operated.

Generally, the two Senior e-Markers spent the majority of their time dealing with items requiring adjudication and items sent for review by markers. In the case of items sent for review, a reply could be sent back to the marker but, as neither the image nor the original query would accompany this, the supervisor had to indicate the question number and make the message very explicit in order to make it meaningful. Initially, both markers and supervisors were not entirely clear whether supervisors should give feedback in response to each message or whether they should simply make a decision on the submitted query and award or change the mark as appropriate. Some confusion may have arisen because in the User Guide it stated rather ambiguously that supervisors could return the item with advice on how to mark it. However, the supervisor had to submit the mark, irrespective of whether they returned it to the marker or not. The home-based supervisor found it sometimes cumbersome to use the messaging system for feedback and expressed a preference for using the telephone. Similarly the centre-based supervisor chose to walk across to talk to her markers and commented that if she were supervising from home she would probably use the telephone rather than the messaging function. Almost all of the supervision in the centre was therefore face-to face.

In practice, both supervisors found that there were far too many items to send back a message every time. Similarly, markers reported that they did not require or expect feedback on all of the items categorised as marking issues, which were sent for review. However, where guidance would have been welcome, as reported earlier, there was no way for markers to prioritise those particular clips so that clarification feedback could be given whilst the marker was still marking the same item. It would probably be helpful to have a distinct category for markers to request 'mark scheme guidance' if this is required. This would probably be even more relevant in a live situation, when more clarification might be required at the beginning of a marking period. Also if a single-marking model is adopted and this system is to be used for feedback/quality assurance purposes it is imperative that the image and original query are returned with the supervisor's message so that markers can assimilate the feedback, annotate their mark scheme where appropriate and adjust their marking accordingly. A large proportion of responses sent to review were BtCIA responses. As it was extremely rare for the supervisors to be able to mark these any more than markers themselves, it was suggested that they should be automatically pulled from the system and not sent to supervisors. This would reduce the workload for supervisors and allow them to focus on marking issues. However one of the major issues

in a scaled-up operation would be making a decision as to how BtCIA responses would be marked.

On a few occasions, it was necessary to send messages to all markers. Although it was possible to request a general message to be added to the Welcome Page that markers would see as they logged on, this had to be done by the technical team. The Chief e-Marker would have liked to be able to add such messages more easily as and when required. A further problem was that the Chief e-Marker found he was unable to add messages to the on-line mark scheme when he was working from home.

Adjudication was necessary when there was a conflict between the marks awarded by the first and second marker of the same pupil response. In the adjudication screen, supervisors were presented with the item on which the markers disagreed and the marks awarded by both markers. The supervisor had to validate the mark s/he considered to be correct or substitute the correct value if s/he disagreed with both markers. However, the adjudication module of the e-PEN system was problematic to navigate around and use efficiently. The drop down menu of items was not in numerical order, resulting in supervisors wasting considerable amounts of time trying to find items on the system. In addition, when adjudicating questions, the system did not automatically bring up the mark scheme for the relevant question. Supervisors had to remember to load the mark scheme separately before starting to adjudicate an item. If they then wished to move to a different question, they had to exit the adjudication mode, change the mark scheme and log back in again. Both supervisors reported that they were able to pick up patterns of errors from particular markers using the adjudication system. In the marking centre the supervisor reported that if she saw a particular marking error during adjudication she would walk across and talk to the marker concerned. The system needs to incorporate this sense of immediacy for the home-based supervisor who does not have this option available. For example, currently messages can not be sent directly from the adjudication screen. The supervisor had to jot down a note of the issue and the name of the marker to send a message later. As the role of adjudication is one of the key duties of an e-marking supervisor (particularly in the double-marking model) these issues need to be resolved as a priority in a scaled-up operation.

In a scaled-up operation, it is possible that several supervisors could be adjudicating items from a large pool of markers and would not necessarily notice common errors from one marker. Recognising errors would be particularly pertinent if the marking model adopted was not 100 per cent double-marked. If this were the case it would be advisable to create small teams of markers whose items would be adjudicated by the same supervisor, in order that errors could be more easily spotted and rapid feedback given. A possible

disadvantage of creating small teams is that inconsistencies between different teams could occur if supervisors make slightly different professional judgements about pupil responses. This undoubtedly occurs to some extent now in conventional marking as during this pilot some small instances of inconsistencies between regions in the interpretation of the mark scheme were revealed. However this possibility could be reduced with e-marking if there was sufficient monitoring of adjudication and review decisions at a higher level.

In the adjudication process supervisors have to decide which marker's response is correct. Thus differences in opinion between markers on items that are difficult to mark have the same status and impact (in terms of reliability) as clear infringements of the marking scheme or careless errors. Depending on the level of accountability, (for example if markers are being monitored or rewarded on the basis of reliability data), in a scaled-up operation this may lead more markers to send items for review, in order to be 'on the safe side', thus potentially increasing the supervision workload.

2.2.6.2 DWS Editor user (data entry) supervision

Although marking within the DWS Editor system was extremely straightforward, making decisions about items sent for adjudication or review of items appeared to be extremely time consuming. It was not always apparent which item was currently on screen, i.e. the paper or tier to which the item belonged. If the supervisor wanted to check the mark scheme (for example to support a professional judgement if the response was somewhat ambiguous), it was necessary to scroll up and down to find the question number and then check the mark scheme for all the different tiers to discover what the correct response should have been. Better on-screen labelling should be implemented to facilitate supervision of data entry items.

2.2.7 E-marker monitoring

As reported above, informal monitoring of the quality of marking was carried out by supervisors by means of the adjudication and back-reading functions. A more formal analysis of the quality of e-marking was provided by NCS Pearson by means of two main monitoring functions. Firstly for all types of markers, an analysis of the double-marking of pupil responses gave a measure of between marker reliability, ie the extent of agreement between a particular marker and all other markers who marked the same item. Expert and clerical markers were also monitored by means of a validity function or absolute-Marker reliability. This determined the extent to which each individual marker was in agreement with the 'absolute' standard set by the Chief e-Marker.

At marker level, reliability data was linked to the first marker and compared against totality of second marks. No inter-rater reliability data was included on a marker if s/he only marked as the second marker, although this was in theory available. For example, one of the home-based markers completed the first marking of an entire item. There was therefore no reliability data on any of the other markers (individually) for that item as they only marked it as second markers. In order to log-on to this report, the supervisor had to enter a threshold percentage for reliability (and validity). Then results would be presented for markers who fell at or below this criterion. There was no guidance for supervisors on what criteria to apply for this report. NCS Pearson have said that in future it will be possible to limit markers to a certain allocation of any one item so that there would be reliability information on each item for each marker. However, consideration should be given as to what reliability data is required at marker level and how this can best be achieved.

Validity items are items that have been marked in advance by the Chief e-Marker (or Examinations Manager) and are intended to provide a check on the quality of marking. Within each 40 clips marked, a marker receives a validity item from a separate pool of such items. Marks awarded for validity items by markers are compared with the 'true' score as awarded by the Chief Marker. The intention is that markers do not know when they are marking validity items and therefore cannot 'sanitise' their marking. Depending on the number of validity items available, once all these have been exhausted the marker will receive the same validity items again. These will be fed in the same order, at a random placement of one validity item within each block of forty items delivered.

Unfortunately the loading of validity items onto the system proved to be somewhat problematic. Initially some prepared items were loaded before the database went live. The intention was that during the marking phase, the Chief e-Marker would examine ('front-read') items from the live database that had not yet been marked and escalate additional appropriate items into the validity pool as required. However, insufficient planning seems to have been given to decisions about how many validity items would be required and the procedures for loading live items into the system. The need to escalate more items appeared to have been overlooked initially. When the Chief e-Marker and Examinations Manager eventually attempted to do this, they found they were unable to do so because of technical difficulties. It later transpired that IT security measures were preventing the images from being transferred into the necessary files. Until this was resolved items that had initially been selected as qualification or calibration items were added to the validity items pool. However, in some cases, it meant that there were no validity items and therefore no validity data available. For other items, there were insufficient validity items relative to the number of items to be marked. As a result,

markers were receiving the same item more than once. A few markers reported recognising repeated items and in one instance a marker commented that he had seen the same item approximately eight times. Another marker recognised validity items from the code in the bottom left of the screen. As it was possible for markers to identify validity items, in effect they were not marked 'blind' and markers may have taken more care on these items. As a result the information gleaned from the validity data will be of very limited use and is a serious shortcoming of the pilot.

The criterion for the distribution of validity clips was set at one validity response for every 40 general responses. The rationale for this particular rate is somewhat unclear. According to NCS Pearson the decision was based on the time available for production, i.e. the more validity responses the more time needed for production. NCS Pearson initially stated that the typical rate at which validity items would be passed through the system would be 10 per cent, (Clip Distribution Rationale for Marking, Version 0.2, 18 July 2002). In their US operations the typical validity rate is 5 per cent, or 1 in 20 responses. The rate used in the 2002 KS3 pilot appeared to be too infrequent to be used for on-going monitoring. Again it is particularly unfortunate that there seems to have been insufficient planning as to what would be required and procedures in place for managing the operation of this function.

Where validity data was produced the interpretation was also not as straightforward as might be at first thought. The way in which validity items were chosen will have a significant impact on how the data should be interpreted. For example, if 'unusual' or 'problem' items were chosen (e.g., unusual interpretations of the mark scheme), then we might expect markers to perform worse than if a representative sample of responses had been chosen. In addition, as reported above, the extent to which the items were marked 'blind' is somewhat questionable. This would mean that performance on the validity items may not give a correct indication of the actual marking standard across all items, i.e., the validity function will not act as a valid quality control check. Table 2.1 illustrates the kind of interpretative problem that was actually faced (and, to a large extent, would be faced on scale-up).

Table 2.1 Extract from marker statistics for question 24b, 2 August 2002.

	no. items read	no. validity item reads	% perfect agreement	no. reliability items read	% perfect agreement
Marker 1	423	11	81.8	862	90.5
Marker 2	472	12	50.0	472	91.3
Marker 3	567	15	53.3	560	90.7
Marker 4	99	3	66.7	98	90.8
Marker 5	234	6	83.3	234	91.9

Note the following points:

- Note that Marker 1 had more reliability reads on 2 August than actual reads, i.e. items marked. This is because there were a large number stacked up from 1 August, which he had marked, but which had not been second marked.

- There were only 6 validity items for this question, which meant that anyone who had marked more than (40x6x2=) 480 items had seen all of the validity feeds twice. The time of presentation of each validity item is randomly allocated within every 40 items. The NCS technician was 'fairly confident' that the software retained session user information such that logging off and on did not affect the order or rate of presentation of the validity items.

- In terms of marker statistics, it is not at all straightforward judging how good is good enough. Marker 2 had a reliability of 91 per cent on 2 August, but only 50 per cent validity. The validity looks poor, but the reliability is no worse than anyone else. This shows that these data are very hard to make good sense of. S/he had 12 validity reads but (at best) only 6 different validity items. It could be that this 50 per cent validity accuracy was actually three items out of six wrong – but the same ones wrong each time. This makes the validity data even harder to interpret.

Much more consideration should be given as to how validity items are to be used and therefore which items should be selected. There should be sufficient items in the validity pool to avoid them being too easily recognised and they should represent a range of responses rather than be viewed as 'training' items. As different questions will be harder or easier than others, it would not seem appropriate to specify a unique criterion of validity accuracy for all items. It is recommended that the e-marking of each question be

piloted (e.g. in a pre-test) before the marking goes live, to determine the appropriate criteria for supervisors to apply on an item by item basis.

The analyses of the formal monitoring of expert and clerical markers was provided by means of a large number of reports available within the e-PEN system. From the supervisors' perspective, information in the Supervisor User Guide about the range of monitoring reports was fairly minimal and was not written in layman's language. There were 18 reports listed in the User Guide, but limited advice as to how and when reports should be used. The Chief e-Marker and home-based Senior e-Marker had used mainly trial and error methods of discovering which reports were most useful. Also there was insufficient written explanation of exactly how the figures in the reports were generated. There was insufficient explanation of the meaning of the reliability and validity reports and therefore what the information available was actually telling them. Overall, the reports appeared to have been designed for post-hoc review rather than to support the day to day supervision of markers. One emerging issue from both the 2001 and 2002 pilots is the capability of e-marking to provide extensive data in respect of markers. However, careful consideration must be given as to how such information should be used. For example, what is an acceptable level of agreement on a validity item? As suggested by one of the Senior e-Markers, in a scaled-up operation e-marking may reveal some unpleasant truths about the level of errors. Provision must be made to tackle this issue with adequate supervision and support for markers.

On the drop down menu of reports there were many more available than those contained in the User Guide, although not all of these were being used within this project. A much smaller list would have been considerably less overwhelming for new users to assimilate. The reports were not particularly user-friendly, as they had to be scrolled horizontally as well as vertically, and column headings disappeared as one scrolled down the pages. Some report formats used item names while others used item numbers, making cross-referencing extremely difficult. Also, clerical items and expert items were listed together rather than separately in many reports. A further limitation of the monitoring data was that information on markers or items was often distributed across several reports. A search facility whereby entering the name of an item or the name of a marker would bring up all the relevant information would have been extremely useful.

One salient anomaly in the data was that of non-adjacent validity errors for single mark items. The validity and reliability reports for an item indicated whether the discrepancy that occurred between two marks was one of high or low adjacency (e.g., a mark of 1 compared to a mark of 2), or non-adjacency (e.g., a mark of 0 compared to a mark of 2). The reports also indicated the direction of this adjacency, showing whether the first mark

was higher or lower than the mark to which it was compared. In some cases, however, reports showed that both high and low non-adjacent discrepancies had been found with items that only afforded a single mark. While it was almost certain that this was due to comparisons involving blanks, the details as to how such results had come about remained somewhat mysterious to the technicians working with the software. They were confident, however, that this glitch could be dealt with in subsequent operations.

The Chief e-Marker found that, whilst at home, downloading times for the reports prevented him from making full use of all the monitoring data. These reports were much easier to access in the marking centre. Also, the reports did not refresh in real time, so supervisors had to close and reopen reports to get more recent data. It was reported by the NCS Pearson project director that the report interface is due to change in the next version of e-PEN, so some of these issues may be resolved at this stage.

In addition to the reports pertaining to reliability and validity, a large volume of other information was generated within the report set. Some of the reports generated by the e-PEN system, if used in a pre test situation, could provide extremely useful information in advance of the live marking phase. For example, the Timing Report could be useful for calculating approximately how long marking would take in a live procedure and possibly, if necessary, how much should be paid per item. However, other reports that contained some useful information could also be potentially misleading. Report I22 gives the number of items sent to review per marker. However, it does not list how many items have been marked to put the review figures into context. Thus one marker may send 50 items to review and another 30, but the former may have marked 600 items and the latter only 150.

NCS Pearson urgently need to address the issue of quality assurance statistics. Rather than the plethora of reports currently provided, it is recommended that there is a streamlining of reports, with consideration as to what information is necessary and how it should be presented to make it both meaningful and useful to supervisors.

2.3 Study 6 – frequency of pupil responses located beyond clip image areas

2.3.1 Introduction

Study 6 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.

This study was essentially a replication of the BtCIA investigation carried out last year in the Evaluation of the 2001 New Technologies Pilot. There were differences in that the sample size was smaller, as described in the methodology section below, and also arising from the format of the KS3 maths papers analysed. These tend not to have specific boxes within which pupils are required to write responses.

Clip image areas were intentionally made larger for this pilot than those in the previous year. Although the clips varied in size according to the item, they tended to extend between one and three centimetres past the perimeter of the space in which a response was expected. In the mental arithmetic papers, clip images were smaller, although a box for the response was provided in this case.

2.3.2 Aims

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.

Note, however, that the study did not directly address the issue as to whether the section of the response that was BtCIA would have had an impact upon the mark awarded to the script. The completed database obtained was sent on to NCS Pearson for an analysis of the impact that responses BtCIA would have had upon marks.

2.3.3 Method

The investigation revolved around the manual scrutiny of scanned full-page images of pupils' scripts. These were inspected using the NCS Pearson Image Viewer. As in the 2001 study, clip image area templates were superimposed over their full-page images, making manual inspection quite straightforward.

These full-page images for each page of the seven answer booklets investigated were accessed remotely on a dedicated website set up by NCS Pearson using a conventional web browser (Internet Explorer). The clip image area template for each item was shown as a light yellow shaded area on the full page image.

On some occasions, the highlighted area for different questions overlapped, meaning that it was not possible to determine where one clip began and the other ended. When this was the case, it was decided to consider the overlapping clips as a single entity. For the sake of

clarity, these are still referred to as 'items' in the analysis, even though they may represent two items with overlapping clip areas.

Fifty scripts were drawn randomly, although without a defined selection protocol, from each of the seven papers investigated in the pilot, making a total of 350 examined scripts. As such, the samples were not chosen to represent the same pupils across papers, as they had been in the 2001 study. The number of scripts was also smaller. 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. Data was thus inputted and recorded directly.

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. As such, the criteria for inclusion of a response BtCIA were exactly the same as those used in the study of the previous year.

Although the examination of the script images and the recording of the data was done by different researchers than those in the study of the previous year, a researcher from the previous year reviewed the work done in its initial stages and confirmed that the same criteria were being applied.

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 in written responses. It also included portions of letters and numbers that extended outside the clip image such that legibility was compromised, for example the tops of numbers such as 6 or 8, or the tops of capital letters such as 'T'. However, it did not include instances where the tips of numbers such as 1 or 4 extended beyond the clip image such that the number in question remained clear. Similarly, instances of numbers such as 2 where the number extended beyond the clip without affecting clarity were also not recorded. Although this was sometimes a strict ruling, and the word or digit 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, to avoid

excessive complication in a binary data set, and to provide data which was entirely comparable with that obtained in the study carried out in 2001.

- Any meaningful annotation that was not necessarily part of the intended answer, per se. This included any working, including working for the mental arithmetic test. The logic here was that mark schemes explicitly required markers to take such annotations into account on occasion.

- 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 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.

After all papers were scrutinised, a 10 per cent quality control check of the entire scripts of 35 students (five per paper) was completed. This resulted in no changes to the database.

2.3.3 Results

Total numbers of responses BtCIA for each item across the sample of 350 scripts and script level prevalence of items BtCIA for each paper are given in Appendix 2. Results of analyses at item level and at script level are described below.

2.3.3.1 Item Level Analysis

Fifty pupils' scripts were evaluated for each of the papers investigated. Prevalence of BtCIA is given by the number of scripts per paper for each 'item' that was affected. Thus an item level prevalence of 50% would mean that half of the responses to that item investigated strayed BtCIA. This analysis therefore investigated patterns between different 'items'. As described above, in some instances, one or more clips were artificially conjoined due to overlap between clip areas. Table 2.2 shows the frequency of responses BtCIA.

As a percentage of all (50) scripts for each 'item', the means of item level prevalence of responses outside the clip area were very similar across scripts (between 5% and 9%). The only exception was the extension paper, which had a mean of 25% of scripts evidencing such responses for each 'item' that was affected.

Maximum percentages of item level prevalence were very similar for Paper 1, levels 3-5 (36%), Paper 2, levels 4-6 (36%), Paper 1, levels 5-7 (38%) and Mental Arithmetic A (36%). Mental Arithmetic C was somewhat lower in its maximum percentage of item level prevalence, with 26%. The Extension Paper had a very high maximum percentage of item level prevalence, with 78%. Paper 2, levels 6-8 also had a high maximum percentage of prevalence at the item level, with 70%. To clarify, this means that one item in this paper occasioned responses that strayed beyond the clip image area in 35 of the 50 scripts examined.

The Extension Paper had the greatest percentage of responses BtCIA on an item of all the papers investigated. Two 'items' had responses outside the clip image area in more than 60% of cases. These were questions 1 (78%) and 3a (64%). Other 'items' with 20% or more of responses BtCIA were 3b (36%), 3c (24%), 3d (28%), 4a,b (34%), 4c (26%), 4e (20%) and 5a,b (20%).

Table 2.2 Frequency of responses BtCIA.

Test	No. of 'items' in this paper	Mean % of responses BtCIA	Test 'items'	
			'Items' with 20-30% of responses BtCIA	'Items' with above 30% of responses BtCIA
Paper 1, level 3-5	42	7%	3c (28%), 6a(i) (20%), 6a(ii) (22%), 16 (24%)	3b (30%), 4 (36%)
Paper 2, level 4-6	35	8%	7 a,b,c (28%), 10b (24%)	15 (36%)
Paper 1, level 5-7	35	7%	4a (22%), 4b (28%)	6c (38%)
Paper 2 level 6-8	32	9%	11a (26%), 11b (20%), 13b (22%)	3 (70%)
Mental Arithmetic A	27	7%	28 (26%)	24 (36%)
Mental Arithmetic C	29	5%	28 (26%)	none
Extension Paper	14	25%	3c (24%), 3d (28%), 4c (26%), 4e (20%), 5a,b (20%)	1 (78%), 3a (64%), 3b (36%), 4a,b (34%)

2.3.3.2 Script level analysis

50 pupils' scripts were evaluated for each of the papers. Prevalence of BtCIA is given by the number of 'items' for each script that was affected. Thus a script level prevalence of 50% would mean that half of the items on a script had responses which strayed BtCIA. This analysis therefore investigated patterns between different scripts, and therefore pupils, within the papers sampled. Again, in some instances, one or more clips were artificially conjoined due to overlap between clip areas.

The mean percentage of items in a paper BtCIA (mean script level prevalence of responses outside the clip area) were very similar across papers (between 5% and 9%). The only exception was the Extension Paper, which had a mean of 25% of 'items' evidencing such responses across each script that was examined. Note that these mean percentages are exactly the same as for the item level analysis.

Maximum percentages of script level prevalence were very similar for Paper 1, levels 3-5 (33%) and Paper 2, levels 4-6 (31%). Paper 1, levels 5-7 (20%) and Paper 2, levels 6-8 (22%) had a somewhat lower maximum percentage of script level prevalence of responses BtCIA.

As a percentage of all 'items' on each paper BtCIA, the maximum percentages were similar for Mental Arithmetic A (59%), Mental Arithmetic C (52%) and the Extension Paper (57%). For these three papers, therefore, there were scripts that had responses BtCIA on more than 50% of the 'items' in the test.

2.3.4 Discussion

2.3.4.1 General comments

All of the papers investigated in this pilot were mathematics test papers. The part of a response found BtCIA was usually pupils' working rather than the actual final answers to questions. This is a potential problem in those cases where marks are to be awarded for working. The questions that did not have as large a number of responses BtCIA were those that provided a box of some sort for recording the answer, rather than a space. Examples of such questions are q 7, Paper 1, 3-5, q 3b, Paper 2, 4-6 and q 9, Paper 1, 5-7.

A high proportion (at an 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 not being seen by the e-marker. This would be because only a portion of a number or word was outside, and the response could easily have been inferred. Although working was noted BtCIA with considerable frequency, only a few questions 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. Working, sketches, graph annotations and other aspects of responses BtCIA for items in the papers analysed should always, in principle, be taken into account, and therefore their visibility is of considerable importance.

In the 2001 study, the clip images superimposed onto all the scripts changed position slightly across scripts. This did not seem to happen this year, and clips always seemed to be in the same place. Central, and therefore well-placed, answers were always captured.

The lack of answer boxes for the students' responses to many items meant that they had little guidance as to the shape or size of the space in which to write their answer. Because they were not aware that their scripts would be viewed and marked as clips, they were sometimes not careful to include all of their writing or working in the designated area. Often, answers were written diagonally across the clip image area, meaning that digits to the extreme left or right of working appeared BtCIA.

2.3.4.2 Paper specific comments

It is of note that the item level prevalence of responses BtCIA for the Mental Arithmetic papers is higher than that in the Mental Arithmetic paper investigated in the previous year's study. Whereas no item in the Mental Arithmetic paper of the Year 7 Progress Test had a proportion of responses BtCIA of more than 10%, in the present study the mean script level prevalence of items that were affected was similar to the main papers (MA A 7%, MA B 5%), with the highest item level prevalence being 36%. Note however that the Mental Arithmetic tests had many scripts with no items BtCIA (42% of scripts for A and 60% for B). This suggests that those students who gave one response BtCIA often made many others, perhaps due to a style of response which involved a lot of working or changed answers.

Although the papers are not easily comparable, and the ability range of the pupils taking them differs, this comparison offers evidence that the mental arithmetic papers are just as much a source of potential responses BtCIA as the other papers. This is despite the limited time given to students to give their answers, and the format of the paper, which gives boxes for responses. It is perhaps relevant, however, that most of the responses BtCIA in this case were for working, which would not affect marks awarded in these papers.

Because of the tiered questions within papers, some items occurred in more than one tier of a paper. This was the case with thirteen item clips which were shared by tiers 3-5 and 5-7 of Paper 1, and 11 item clips which were shared between tier 4-6 and tier 6-8 of Paper 2. Few obvious trends can be discerned in the data for these items. For some common items there were more instances of BtCIA in the lower tier and in others there were more in the higher tier. One shared item which is of interest is question 3 of Paper 2, tier 6-8, which also occurs as question 15 on Paper 2, tier 4-6. The incidence of responses BtCIA in this item in tier 4-6 is approximately half (36% as opposed to 70%) that found in tier 6-8. The candidates in Paper 2, tier 6-8 appear to have written more working, which could account for the increased incidence of responses BtCIA. This clip image area is not wide

enough to accommodate the two columns of workings which candidates often produced, resulting in responses extending beyond the right boundary of the clip image.

The prevalence of responses on the Extension Paper is noticeably higher than that on the other papers investigated. Script level prevalence for this paper was as high as 57% with a mean of 25%, and in the case of one 'item' (question 1), 78% of all respondents gave a response BtCIA. It seems likely that this high rate of response outside the clip image area is due at least in part to the high ability range of students taking this test. As such, they were perhaps able and willing to demonstrate their knowledge to a greater extent than students in other papers, including a great deal of working and evidence for many items.

In the case of question 1 on the Extension Paper, the clip extended across half of the page, and a lot of working was required. For this reason, if mistakes had been made and new working had to be started, it is likely that responses would have extended BtCIA. Also, this question began with a diagram, and many annotations were made on this. The same points are true for question 3a of this paper (64% BtCIA), which is a very similar item. Question 3a also requires that a formula is applied.

One item noted in particular as problematic by the researcher recording instances of responses BtCIA was question 4 of Paper 1, levels 3-5. This question was the only one presented on the page, and as such the clip area was quite large (approx. ¾ of an A4 test page). Many of the students provided their answer to this question at the end of the line on which the question was presented, the problem with this being that the questions themselves are not included in the clip image. As the clip area started just below the last line of the question, students who failed to provide working and only gave the answer in this space would not have had their answer viewed by a marker. This problem might be overcome by extending the clip image area to include the last line of the question, or by providing a box in which both working and answer could be provided. This would also have helped those candidates who made errors in their working and began the next set of working in a space adjacent to the previous set, again just outside the clip area.

Another type of potentially problematic item was one, which was accompanied by data, such as a graph or table that was necessary to answer one or more parts of a question. In these instances, the graph would be presented before the question, and as such would not be included in the clip image area. This meant that if a candidate made any annotations to this data (a particularly common practice was to show working on a graph) they would not be seen by the marker. This constitutes a source of unreliability in marking in itself, and it also potentially affects the accuracy of the results of the present study, because anything that was written on the diagrams was counted as BtCIA of the first part of the question.

This is potentially misleading because the annotations sometimes referred to all or to a different part of the particular question.

One problem which arose as a result of having larger clip image areas was that, in cases where more than one question appeared on a page, the clip image areas would sometimes overlap. Because these areas were not uniform, it was sometimes difficult to say where one clip ended and the other began. This meant that, although a student's answer might appear to be inside the clip image area, it could actually have been extending into the clip image area of the next question, and as such be BtCIA. For this reason, during initial checking of the clip image areas on each paper, these clips were classified as 'pseudo-conjoined clips'. This problem occurred to some extent on all of the papers that were evaluated. Although many questions are presented on one page in both of the Mental Arithmetic papers, the clip images on these papers were a more uniform size (approx. 1cm x 3.5cm), and it was therefore easier to estimate where one clip image area ended and the next began. This meant that not so many had to be classed as psuedo-conjoined clips.

2.3.5 Summary

The frequency of pupil responses which extend outside the clip image area is considerable. With the exception of Mental Arithmetic C, each paper evaluated has one or more items in which the incidence of responses BtCIA was greater than a third of those pupils taking the test. In the case of the Extension Paper, a mean percentage of 25% of responses were BtCIA across scripts. Although it is salient that many of these responses are not seriously compromised in their legibility, or do not afford information that may merit a change in the e-markers' evaluation in the mark to be awarded, there can be no doubt that with such a high rate of incidence there are a significant number of responses in which these issues are relevant.

As such, the problem of responses or annotations outside the clip image area remains a source of threat to e-marking validity in a scaled-up operation. The increase in size of the clip area from that used in the BtCIA study in 2001 has not had the desired effect of reducing the incidence of responses BtCIA found in a paper. It is relevant, however, that the different tests investigated in this study and the pupil sample are not directly comparable with those studied in the previous year's analysis.

As found in the 2001 study, responses BtCIA are usually due to rough working. Across all papers, some pupils have provided answers within the clip image which are subsequently crossed out. The new answers are often written outside the clip area, sometimes entirely outside the view of an e-marker. In some cases the clip image area has not been sufficiently extended to include areas where students often make annotations,

either in or after the text of the question itself, or, more commonly, on the graphs or data tables provided.

It seems that a possible solution to the problem of responses BtCIA would be to delineate the area where students are expected to give their response clearly, with a box rather than a space. Clip images should be as large as possible, and should extend into areas other than that given for the response per se, such as the question's graphs and tables, to increase the chance of capturing students' working. Further, it is to be expected that students will give fewer responses BtCIA if they are made aware during the administration of a test that their work is to be scanned and marked as a clip, especially if the borders of this clip are marked on the paper.

2.4 Summary of key issues arising

In general in this pilot, NCS Pearson demonstrated (albeit on a small scale) that the e-marking of KS3 mathematics test papers can be undertaken using either a centre-based or home -based model of on-line marking. However, this was not without problems. The markers and supervisors participating in the pilot were generally extremely positive about the e-PEN system and the-benefits of the e-marking process. Whilst recognising the extent to which NCS Pearson were largely successfully in implementing the agreed procedures for 2002, the following section will attempt to summarise the most significant shortcomings of the pilot and the implications for any scaled-up operation.

- the failure to scan at the forecast rate of over 5000 documents per hour;
- the failure to deliver the practice, calibration and qualification functions;
- the failure to fully test the system prior to the live phase and therefore anticipate some the hardware/software problems for home-based e-Markers, particularly in respect of the size of the image files;
- the failure to fully test the system prior to the live phase and therefore anticipate some of the frustrating peculiarities of the software that caused problems for markers (e.g. zoom and overlay functions) and supervisors (e.g. adjudication mode);
- the failure to anticipate the cross-referencing problems within the report set and the failure to provide sufficiently tailored monitoring data to support the supervision process;

- the failure to put in place adequate procedures for monitoring validity (absolute-Marker reliability) with the consequent loss of data;
- the failure to provide adequate training and comprehensive User Guides for supervisors.

It is important to note that these points are expressed as shortcomings of the pilot. Some of the failings listed above are solely the responsibility of NCS Pearson for failing to deliver system requirements that were promised. In other instances culpability may not lie entirely with NCS Pearson. It is probable that some of these shortcomings could have been avoided had there been more time and attention devoted to the planning and specification phases of the project.

3 Questionnaire feedback from e-markers

3.1 Introduction

This section of the report describes the findings from the exit questionnaire given to both the home-based and centre-based markers at the end of the marking period. The methodology is described in section 3.2, followed by discussion of the outcomes in each main content area covered in the questionnaire. These are: training (section 3.3); software/technical/marking issues (section 3.4); technical support (section 3.5); marker supervision (section 3.6); expectations and experience of e-marking (section 3.7); and views on scaling-up the system (section 3.8). Qualitative comments are discussed in section 3.9 and the questionnaire findings are summarised in section 3.10. A copy of the exit questionnaire and the full data from Study 3 are presented in Appendix 3.

3.2 Methodology

3.2.1 Method

A single exit questionnaire was developed for both home and centre-based markers. The questionnaire asked markers to express their views on the success of the 2002 new technologies pilot and, more generally, on the potential for scaling-up similar systems and procedures to a national level for key stage 3 maths. Markers were also asked to provide background data to facilitate analysis. This included details of previous marking and teaching experience as well as personal information, such as gender and age. The questionnaire items were predominantly closed response, to facilitate both completion and analysis. A space was also included for additional comments, and these are referred to in section 3.9 below.

All of the 21 home-based markers and 16 out of the 17 centre-based markers returned questionnaires. Although three of the home-based markers were also involved in centre-based marking, they were asked to complete the questionnaire drawing upon their experiences as home-based markers.

3.2.2 Sample

Background data for the markers returning the questionnaire was as follows:

- Gender: 65 per cent of the sample was male and 35 per cent female.
- Age: Markers gave their age in years. Seventy-five per cent of the sample were aged between 40 and 60. Appendix 3 shows responses in 10-year bands.

- Current job status: Of the 37 markers who completed the exit questionnaire, 15 were full-time teachers, three were supply teachers, two were LEA advisers, one was an educational consultant, 12 were retired and one was unemployed.

- Teaching experience: All of the markers had been employed as teachers. Three of the markers had most recently been employed as a classroom teacher before 1995, 10 between 1995-1998 and 24 from 1999 to the present.

- Marking capacity: In 2002, three markers had been appointed as senior markers for conventional marking, 18 had been appointed as team leaders and 16 as markers.

- Marking experience: Only three of the markers had not been appointed as external markers for national curriculum tests before 2002. Twenty-five of the markers had been employed as external markers for the past seven years.

A recent survey of external markers by the NFER gave details of the AQA marking population across key stages 2 and 3 and enables a comparison of the profile of the pilot sample with that of a large and representative population of markers in England. The sample of markers in this pilot was more predominantly male than the national average, which has only 37 per cent male markers. It is also clear that the pilot sample involved a considerably larger proportion of senior markers and team leaders than that of the population of markers as a whole. This has senior markers as two per cent of markers, and team leaders as making up only 10 per cent. The sample population in the 2002 pilot had eight and 49 per cent respectively. Although the data in respect of AQA markers is not specific to key stage 3 mathematics, it is likely that these differences do reflect aspects of the pilot sample population that vary from the average. It is possible, however, that this comparison performed in relation to a representative population might be unfair, in that the key stage 3 mathematics marker population might be different from that for other subjects.

It is important to bear in mind that the total sample of 37 markers is a small number. Analysis of the data below can only identify areas that may be of significance in a scaled-up system, rather than to draw certain conclusions.

It should also be remembered that percentages of the small numbers involved can be misleading, although these are included in the discussion below for information.

3.3 Training

This section of the questionnaire (section 1.1) asked markers whether they considered that the amount of training which they had received on the e-PEN software was sufficient, and

the extent to which they were confident initially using this software to score responses and receive messages. Markers were also asked how they would feel about receiving e-marker training entirely on-screen, without the presence of a supervisor.

All but two of the markers considered that the amount of training which they had received had been 'about right'. Both of these markers graded the amount of training as 'too little', and both of these were home markers.

All but two of the markers graded themselves as being between 'fairly confident' and 'very confident' when they began to score student responses. There was a much wider spread of responses on the issue of initial confidence in receiving messages using the e-PEN software, however. Here approximately a quarter of markers felt themselves to be below 'fairly confident', the mid-point of the scale.

Markers tended to be against the idea of training being provided entirely on-screen, without a supervisor being present. Across both home and centre-based markers, less than a third considered such a proposal to be satisfactory. Almost three quarters of respondents were either undecided or said that they would feel dissatisfied with such an arrangement.

3.4 Software/technical issues

Results discussed in this section come from the section of the questionnaire titled 'Marking' (section 1.2). Here markers reported on their experience of using the e-PEN software, the amount of time it took to score one response, and their reactions to marking only a limited number of questions.

Responses showed that markers generally found the e-PEN software amenable to understanding and practical use. Only one marker rated the system/interface as 'difficult' to use, while more than three-quarters rated it as 'easy' or 'very easy' to use. There was quite a marked difference between home- and centre-based markers on this point, however: whereas a third of home markers rated the software interface at the midpoint of 'OK' to use, 100 per cent of centre markers rated it as 'easy' or 'very easy'. This gives an indication that the support available in the centre, even if this was simply one afforded by a shared experience, made a significant contribution to markers' confidence with the e-PEN interface.

Responses to one question, which asked markers to estimate the average time taken (in seconds) between submitting a score and the system allowing the next item's mark to be entered, clearly show one of the main differences between the experiences of home and centre-based marking. Eighty per cent of home markers estimated this time as between

nine and 30 seconds, with an average of approximately 14.5 seconds, while centre-based markers' estimates had an average of approximately five seconds. An analysis of the waiting times experienced by home markers by their location (urban, suburban or rural) suggests that rural home markers had longer waiting times: their mean reported wait was approximately 20 seconds, while the suburban mean was approximately 14 seconds.

This longer submission period reported by home markers may have important implications for the practical marking of many items. It is also possible that this extended waiting time may affect markers' concentration, or ability to get into an efficient 'rhythm'. While the great majority of both groups of markers said that they only considered giving up e-marking because of the speed of the connection 'occasionally' or 'never', in the centre marker group only two chose 'occasionally', while in the home marker group this category was chosen by nine markers (45% of the group). Further, two home markers stopped marking 'frequently' or 'very frequently' for this reason.

Home markers also reported that it was more common for them to be 'frozen out' of the system, with a third of them describing this as happening 'fairly often' or more frequently. All but one centre marker described this occurrence as happening 'occasionally'. It is likely that the range in experience between home markers is caused in part by the range in the specification of their computer hardware. It is of note, however, that four of the home markers (20%) reported that they got frustrated with the e-PEN interface 'fairly often' or more frequently, while no centre marker chose these categories.

Questions also investigated markers' opinions on marking one item for a long period of time. Estimates of average times spent on the same question showed similar results between groups, with most markers spending between 30 minutes and two hours on one question. In both groups, opinions were spread quite equally between 'yes', 'no' and 'depends' for the question 'would you mind having to mark the same question for a whole day?' Most markers described themselves as between 'undecided' and 'very satisfied' with the prospect of only marking a limited amount of items.

Most markers felt that on-screen marking made them feel 'neither more nor less professional' than conventional marking. Of the eight markers who did not choose this category, four centre markers felt more professional, while three home markers felt less professional.

3.5 Technical support

This section (section 1.4) asked markers about the times when they were in need of, and their level of satisfaction with, technical support.

Of the home markers, five (24%) reported needing technical support outside the hours of 8:00 and 23:00. This reflects the wide range of times which home markers chose to mark. Out of all the markers, only one home marker reported being dissatisfied with the level of technical support provided. Three quarters of markers were satisfied with the level of this support, while the rest were either undecided or had received no such support.

3.6 Supervision

This section (section 1.3) asked markers about the 'send to review' function and the effectiveness of supervision by e-mail, and compared the level of supervision to that received during conventional marking. The questionnaire also investigated opinions as to which aspects of the feedback received were most appreciated.

The majority of markers (86%) agreed that the 'send to review' function was 'quite important' or 'very important'. While centre markers reported that 'send to review' items sent to a supervisor almost never received a response, three quarters of home makers received such a message 'occasionally' or more frequently. Approximately a half of home markers were 'satisfied' or 'very satisfied' with the speed of these responses, while a quarter were 'dissatisfied' or 'very dissatisfied'.

Only four markers, all of them home markers, received individualised supervision on a personal (non-e-PEN) e-mail account. This guidance was generally thought to be useful.

There was quite a wide spread of opinion between markers as to how preferable 'real-time' interactive supervision was, compared to electronic message supervision. Centre markers tended to think that 'real-time' supervision was more often preferable than home markers did, although the difference was small. Fifteen home markers (71%) felt that they received less supervision than they had during conventional marking, and even in the centre, seven markers (44%) felt that the level of supervisory guidance was similarly less. Most other markers felt that supervision was about the same.

Approximately 60 per cent of markers would have appreciated more feedback as to the speed of their marking in absolute terms (e.g., average number of items per minute), while approximately a third were undecided. Fewer markers expressed a wish for more feedback as to the speed of their marking in terms of their relative standing to other markers, although this was true of approximately 40 per cent of markers in both groups. Such

feedback was less welcomed by centre-based than home markers however, with six centre markers saying that they would not appreciate more of such feedback, compared to only three home markers who reported feeling this way. This perhaps reflects centre markers as having more experience of working in a team to mark, in which case such relative feedback might be damaging to team spirit.

When markers considered more feedback as to the accuracy of their marking, the large majority of both groups of markers reported that they would appreciate more information as to their absolute accuracy (percentage of marks validated). More than a half of both groups also felt that they would have appreciated more feedback as to their relative standing in terms of accuracy.

3.7 Expectations and experience of e-marking

This section (section 2) asked markers about their expectations before starting e-marking, and their actual experience of several aspects: the level of intellectual demand, the speed of marking, and the levels of satisfaction, stress, interest and fatigue compared to conventional marking.

In retrospect, the markers' expectations of the e-marking pilot were reported as follows:

- Intellectual demand: the great majority of markers expected this to be about the same, or had no prior expectation.
- Speed of marking: Slightly more than half of the markers expected e-marking to be faster than conventional marking. No marker expected it to be slower.
- Satisfaction with marking: Most markers expected this to be about the same, or had no prior expectation.
- Level of stress: While most markers expected this to be about the same or had no prior expectation, approximately a quarter of markers reported that they thought e-marking would be less stressful than conventional marking.
- Level of interest: Overall, approximately half of the markers thought this would be about the same or had no expectations. Most other expectations were that e-marking would be more interesting. Somewhat more of the centre-based markers expected a more interesting experience than the home-based markers.
- Level of fatigue: Approximately half of the markers expected e-marking to be less tiring or about the same as conventional marking. Approximately a third had no

expectation. Approximately a quarter of the centre-based markers expected e-marking to be more tiring.

Then, markers reflected on their actual experience of e-marking in terms of the aspects outlined above:

- Intellectual demand: The majority of centre-based markers reported that this was about the same. Approximately a third of home markers reported that e-marking was less intellectually demanding than conventional marking.

- Speed: Here again, a clear distinction between home and centre markers was shown: while approximately 70 per cent of centre markers found e-marking faster than conventional marking, just over half of the home markers reported that they had found it slower.

- Satisfaction: Views were quite widely spread between markers on this point. Home markers tended to rate e-marking as less satisfying than conventional marking, while centre markers tended to report that it was more satisfying. A cross-tabulation analysis suggested that markers with team leader status tended to rate e-marking as less satisfying, while those who were simply markers tended to rate e-marking as more satisfying. While 62.5 per cent of markers rated e-marking as 'more satisfying', only 16.7 per cent of team leaders rated it in this way.

- Stress: The majority of both groups of markers rated their experience of e-marking as about the same or less stressful than conventional marking.

- Interest: While home markers were evenly spread as to whether e-marking was more or less interesting than conventional marking, more than half of the centre markers reported that it was more interesting. This may have been due to the social aspect of working with other markers.

- Fatigue: Many markers reported e-marking as being about as tiring as conventional marking. Other views were balanced quite equally between more and less tiring across both groups, with less tiring being chosen by slightly more markers in both cases.

Analysis of whether markers had changed their view from their initial expectations to their actual experience showed that most views had remained the same. Those views that had reversed, however, were always towards the negative: seven home markers found that e-marking was slower than they had expected, and of these, four had reversed their view from faster to slower. Two home markers who had thought that e-marking would be more interesting than conventional marking found that it was less so. Similarly, one

home marker who had expected e-marking to be less tiring than conventional marking found that it was more tiring.

3.8 Views on scaling-up the system

This section (section 3) asked markers to give their opinions on various implications of scaling-up the e-marking system. Aspects specifically investigated included: accuracy, the impact upon the construction of the key stage 3 maths test, effectiveness of supervision, time taken to mark, marker preference, teacher acceptance, and marker thoroughness.

- Accuracy: Almost all markers from both groups were either unsure or agreed that e-marking would be likely to be more accurate than conventional marking.

- Future test construction: Similarly, almost all markers from both groups were either unsure or disagreed that e-marking would have a negative impact upon the construction of future key stage 3 maths tests.

- Effectiveness of supervision: The large majority of markers thought that supervision of on-screen marking is likely to be more effective than supervision of conventional marking. No markers from either group disagreed that this would be the case.

- Marking time: More than half of markers were undecided as to whether on-screen marking would be likely to make the overall marking period shorter than for conventional marking. Perhaps not surprisingly, of those who expressed an opinion in either direction, most of the home-based markers thought that e-marking would take longer, while the majority of centre-based markers thought that the new system would reduce marking times.

- Preference: Results here reflected those for the perceived marking time. Approximately a half of markers were unsure whether markers would prefer the new system. Approximately 40 per cent of home markers disagreed that markers would prefer e-marking, while about 40 per cent of centre markers thought that markers would, in principle, be likely to prefer on-screen marking to conventional marking. A cross-tabulation analysis indicated a tendency for markers to predict a preference for e-marking, while team leaders were more likely to predict that markers would not be likely to prefer the new system.

- Teacher acceptance: Just less than half of the markers were unsure as to whether most key stage 3 teachers were likely to accept on-screen marking. Most others felt that it would be accepted. This feeling was similar across both groups.

- Careless mistakes and thoroughness: There was a slight discrepancy in results to a question which asked markers whether they felt that markers would be less likely to make careless mistakes when marking on-screen and one which asked whether on-screen markers would be more likely to be thorough. These questions would seem to be very similar in their focus, yet results showed that approximately a third of markers from both groups disagreed with the first statement, while just under a half of markers from both groups agreed with the latter. In both cases, the majority of other responses were for a 'don't know' response. Perhaps this apparent anomaly was due to the structure of the questions, which involved some potentially tricky double negative statements. It may, however, be the perception among markers that e-marking tends to lead both to more thoroughness, but also to more careless mistakes. It is also possible that markers' perceptions could have been affected by knowing that a double marking model was being applied.

3.9 Other comments

This section reports those additional comments which were included by markers in a written section of the questionnaire. Comments were coded, and those which were made by more than one marker are reported below, in the following sub-sections: limitations, benefits, recommended adaptations and other. More comments were made by home markers on the questionnaires, perhaps because centre-based markers felt that they had had more opportunities for passing on feedback.

3.9.1 Limitations

Two markers mentioned the problem of portions of students' answers being out of clip. Four markers reported having problems with using the overlay function of the software, and its rotation. Five markers, all of them home markers, mentioned delay between the delivery of items. Similarly, two home markers felt that marking on-screen felt slower than conventional marking. Six markers mentioned frustration at being unable to correct errors once a mark had been submitted. Two reported that they felt less confident with their marking accuracy when marking on-screen.

3.9.2 Benefits

Seven markers, representing both marker groups, felt that on-screen marking was a successful system. Two remarked positively on marking a restricted set of questions, and felt that this would improve accuracy. Two felt that double marking would improve accuracy, and two commented positively on the training received. Two markers reported

that supervisory support had been good, and three markers noted the benefits of having administrative tasks eliminated by the new system.

3.9.3 Recommended adaptations

By far the most common suggested change, made by five markers in each of the groups, was for a function whereby the most recently submitted item could be recalled to amend errors. Two markers suggested having access to the whole page in cases where responses went outside the clip area. Two markers felt that students should be provided with answer boxes to lessen the occurrence of responses outside the clip.

Two markers felt that it would have been beneficial to have a series of test items before each item was marked. Three markers, all of them home-based markers, suggested having a sound to alert them that a new response had arrived. Three markers suggested having an on-screen tally to show how many items had been marked.

3.9.4 Other

Three centre-based markers mentioned that working in a centre meant that they could more easily get help from colleagues and supervisors than conventional marking would have afforded.

3.10 Summary

The most salient conclusion to be drawn from the exit questionnaire of e-markers is a noticeable difference in the experiences of home-based and centre-based markers. While both groups of markers tended to be positive about the benefits of on-screen marking, home-based markers were somewhat more likely to note negative aspects of the system.

Written comments given by markers again showed that markers overall tended to be positive about the new system, with seven markers making positive comments about the system, while none directly criticised it. The major limitations described were the delay in receiving items, and problems with some items being outside the clip image area. Markers were also frustrated at being unable to recall an item to which they had given an erroneous mark once it had been submitted. A function whereby such recall would be possible was the adaptation most often suggested by markers.

Although the sample size was small, and statistical significance could not be calculated with confidence, it seems that the most pressing issue in the markers experience of e-marking is that of the time in receiving items. This was a much more significant issue for those working at home and had an adverse effect on their experience of the process of on-

screen marking, and it is this which should be principally borne in mind as in need of development when a scaled-up operation is considered.

4 A statistical analysis of the pilot

A major component of the 2002 evaluation was a statistical analysis of the data arising from the pilot. Study 4 collected quantitative data in two categories: management data (information on the speed and accuracy of procedures for the processing of scripts through the electronic marking system); and measurement data (information on the accuracy of results arising from the e-marking). Study 5 was a small-scale, controlled evaluation of marking reliability that allowed a direct comparison between conventional and new technology marking. The results of Study 4 are presented in two separate sections, 4.1 and 4.2, corresponding to the management and measurement data respectively. This is followed by a report on the findings from Study 5 in section 4.3.

Study 4 began with the preparation of *A Framework for Evaluating the New Technologies Pilot 2002*, developed by the NFER in collaboration with NCS Pearson and QCA. This outlined the management and measurement data that would be required to complete the statistical analyses.

Although only one subject area was explored in the 2002 pilot, compared to both English and mathematics in 2001, the analysis of data was quite extensive. Key stage 3 mathematics has a number of discrete tiers, which were analysed separately. Thus there were four tiers for each of Papers 1 and 2, plus three Mental Arithmetic Papers and an Extension Paper. In addition, analyses of the expert markers had to be replicated for home- and centre-based markers

Initial checking of the data supplied by NCS Pearson revealed several omissions and anomalies. Aspects of both the management and the measurement data had to be rejected several times before the data supplied was found to be satisfactory for further analysis. Given that, in the short time allowed for verification of the data, several errors were discovered, it is entirely possible that other errors remain undisclosed. In some cases, it is difficult to be fully confident about what remains and unfortunately this needs to be borne in mind when interpreting the data that follows. In a scaled-up operation, such a level of error would be unacceptable, particularly if the data were being used to evaluate marking reliability. It is essential that, in any scale-up of this pilot, the data production systems be carefully monitored, to ensure that the data is both accurate and meaningful, and that those producing the data are able to evaluate the nature of the data being produced. Specific issues are discussed in the appropriate sections.

4.1 Analysis of the management data

The first major component of Study 4 was the examination of management data in respect of the following:

- time taken to process scripts through the batching, splitting and scanning stages;
- time taken to mark individual questions;
- supervisory and adjudication demands;
- prevalence of various processing errors.

The central focus of the analysis of the management data was whether the implementation of new technology marking could be expected to eliminate 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).

A further purpose of the 2002 evaluation was to determine whether the findings of the 2001 pilot would be replicated in a new context. Wherever possible, findings from 2002 are therefore compared with those of the previous year. However, not all data could be compared directly, as the requirements of the 2002 pilot differed in many respects from that of 2001. For example, in 2002 there was no scanning of exceptions, attachments or full page images.

The data underlying the following analyses are presented in Appendix 4.1. They were provided for the NFER by NCS Pearson. (The tables are not exactly as provided, as a number of presentational and statistical modifications have been made.)

4.1.1 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.

4.1.1.1 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 split/guillotined and scanned into the system?***

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

Table 4.1.2A[7] (Appendix 4.1) presents information on the time lag between the end of processing of a batch of scripts on DWS Batch Builder (i.e. the logging in of scripts following their receipt from schools) and the commencement of processing of the same batch on DWS Scan Master. This includes time taken to guillotine/split each script within a batch (i.e. the removal of the spines prior to scanning).

As can be seen from the standard deviations in this table, there was considerable variation in time lag across batches (for each of the twelve papers). This was supported by an analysis of the minimum and maximum time lags: across papers the smallest lags ranged between one minute and one hour and 54 minutes, while the largest lags ranged between 3.5 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 44.3 hours for Mental Arithmetic A to 67.2 hours for the Extension Paper. These means are considerably longer than the equivalent figures for the Year 7 mathematics progress tests in the 2001 pilot, when the mean time lags ranged from 19.3 hours (maths B) to 22.1 hours (maths A).

The principal cause of delay in processing the batches occurred when scripts could not be scanned, due to a range of problems (see 4.1.1.2 for a more detailed discussion). Significantly, even if there was only one 'rogue' sheet in a batch, the entire batch was held up while the script was removed. As has already been reported (Section 2.2.3), a large number of exceptions would not be desirable in a scaled-up operation.

Table 4.1.4A indicates that the number of 'rogue' scripts ranged from 38 to 680 across papers. For the main 11 papers, the percentage of 'rogue' scripts ranged from 6.5 per cent

[7] The suffix A (as in Table 4.1.2A) indicates that it is located in the Appendices.

(Mental A) to 19.3 per cent (Paper 1, Tier 6-8), with the Extension Paper being something of an exceptional case at 36 per cent. Evidence from the BtCIA study (section 2.3) indicates that, on a small sample, 25 per cent of responses on the Extension Paper were outside the clip image area. As can be seen from Table 4.1.4A, the percentage of scripts that were not fully scanned increased in the higher tier papers. This may be due to students in the higher tiers including more working out, that obscured the pre-printed 'timing marks' added to enable scanning. Overall, approximately 41,000 scripts were received during Cycle 1, of which 10.5 per cent were not fully scanned.

Table 4.1.5aA shows the scanning frequency of student script sets that were returned in Cycle 2, the marking phase of the pilot. Each set of scripts should comprise Papers 1 and 2 of an appropriate tier, plus one of the three mental arithmetic papers (with an optional extension paper for students taking the highest tier). Although 5,642 valid script sets were received back from schools, only 4,213 of these (approximately 75 per cent) had been fully scanned during Cycle 1.

Table 4.1.6A presents information on scanning rates for each of the twelve papers. It shows that average scanning rates ranged between 547 sheets per hour (Extension Paper) and 2,819 sheets per hour (Paper 2, Tier 3-5). The quickest scanning rates ranged between 1,449 sheets per hour (Extension Paper) and 6,114 sheets per hour (Paper 2, Tier 3-5). Across all the papers, the average throughput of the pilot was approximately 2,200 sheets per hour. This was slower than the average scanning rate of 2,880 sheets per hour achieved during the 2001 pilot. The way the performance of the scanners has been reported is to take the time from the start to the end of scanning and to calculate a sheets-per-hour figure based on the number of complete documents processed. In other words the sheet count does not include any sheets from partly scanned documents; those that were scanned but then removed because an exception occurred.

At the quickest scan rate, only one paper exceeded the rate of 5,760 sheets per hour used by NCS Pearson for planning purposes. In the 2001 pilot, the main reason given for the slow scanning rates was the need to scan each script for both a full-page image and for item clip images. This year, as outlined in section 2.2.3, the main reason given initially for the slow scanning rates was the extremely high number of exceptions. However, this in itself does not explain the differences in scanning rates. For example, the scanning rates for the mental arithmetic papers were low and yet they did not suffer particularly high levels of exceptions. The nature of the contribution of the different variables; exceptions, number and size of clips, scanning resolution, etc. to the slow scanning rates was not clear from the management data. Subsequently, further information from NCS Pearson revealed several factors, specifically related to the pilot, which they felt had also

contributed to the slow scanner performance. First, the productivity of the temporary scanner operators, employed for the duration of the pilot, was less than that of experienced NCS Pearson employees. Second, when an exception occurred the scanner operator had to remove the document and communicate the reason for the scanning failure to a clerk who was completing a scanner exception log. In some cases, it appears that this procedure led to delays in re-starting the scanning. Third, scanner performance was adversely affected by batches with relatively low numbers of sheets (for example, the mental arithmetic papers) in that the time delay at the start and end of a batch was spread across fewer sheets.

As there was no requirement to scan the exceptions in the 2002 pilot, it is not possible to estimate the further delays that would have resulted if it had been necessary to scan the exceptions and attachment sheets on a flat-bed scanner. Although the frequency of scanning errors alone might constitute a threat to the timely processing of test scripts, the level of exceptions was higher than would be expected in a full scale-up, given that such test papers would be designed to prevent or minimise the problem of students writing in the wrong place (see 4.1.1.2, below). However, a true picture of the scanning rates that could be achieved may only emerge if a test paper that was designed to optimise scanning is piloted in a sufficiently large quantity.

Tables 4.1.1aA and 4.1.3A present summary information in respect of Cycle 1. Table 4.1.3A gives information on the time lag between the logging of a batch of scripts into the system via DWS Batch Builder and the despatch of the same batch to a marker via DWS Mail Manager. This includes the time taken to split and scan the test papers. On average, the time lag between post-receipt batch processing and batches being ready for despatch ranged from 58.3 hours for Mental Arithmetic A to 80.7 hours for the Extension paper. This compares with the target turnaround of 48 hours.

4.1.1.2 Study 7 – Scanning exceptions

Question: ***What were the main reasons for papers failing to scan?***

In collaboration with NCS Pearson, the attachments and scanning exceptions recorded during Cycle 1 were coded and analysed.

Each attachment or scanning exception was recorded by NCS Pearson on a data collection form. On each form, information was supplied in respect of the barcode, the test paper, the scanner operator ID and the reason for the attachment or exception. Open responses were coded at NFER and data analyses were carried out.

Initially, the database consisted of 4,378 instances of either an attachment or an exception. However, an examination of the barcode file revealed that there were 268 duplicates. Enquiries at NCS Pearson identified several reasons for the data collection forms having been duplicated. In the majority of cases (over 200), duplicates had been created when the forms were being photocopied for despatch to the NFER. A cross tabulation of the reasons given for duplication with the scanning operator IDs revealed that almost all the photocopying errors were due to one operator. The duplicated forms were removed from the database, resulting in an analysis of 4110 cases. Data on Table 4.1.4A from NCS Pearson shows that 4,264 scripts were not fully scanned, indicating that there were approximately 150 attachments or exceptions for which no data collection forms were completed. Of the 4,110 cases, there were 55 instances of attachments and 4,055 instances of scanning exceptions. Of the attachments, the vast majority were due to the student making use of an amanuensis.

The main reasons indicated on the data collection forms for scanning exceptions are shown in Table 4.1.

Table 4.1 Reasons for scanning exceptions

	Number	Percentage of exceptions
Pupil damage to barcode	44	1.1
Pupil damage to Page ID marks	4	0.1
Pupil damage to timing mark	288	7.0
Pupil answer into timing mark	3333	81.1
Poor guillotining during print	2	0.0
Poor slitting prior to scanning	19	0.5
Poor printing	105	2.6
Paper damaged	89	2.2
Other	150	3.6
No response	86	2.1
	4120	

As can be seen from the above table, by far the most common reason for pages failing to scan was because student responses had extended into the timing marks. This suggests that if the area available for the student's response could be clearly defined within the test paper design, and students could be encouraged within the administration rubric to write only within the specified area, most of the scanning problems could be resolved.

4.1.2 Marking of scripts

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

Turning to the marking of scripts, the software ostensibly 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.

Tables 4.2.1A, 4.2.2aA and 4.2.2bA present productivity data for clerical markers, centre-based expert markers and home-based expert markers, respectively. The total elapsed time includes all scoring activity for that item, including first and second marks, scoring of validity items, reviews, adjudication, etc. The mean time per clip has then been calculated for each item. Where clips are conjoined (i.e. two or more items scanned together as one clip) the mean time is given for marking the clip as a whole. Data from these tables are summarised below, in Table 4.2.

One minor technical caveat applies to the productivity data and also to all the management data and measurement data relating to marking. During the e-marking, three items originally categorised as expert items were partly marked by clerical markers. However, no data relating to these items appears within the relevant tables for clerical markers. On querying this anomaly with NCS Pearson, the explanation given was that all the data had been included within the expert data, as this was how the items had originally been classified within the framework document. Asked to quantify this, NCS Pearson reported that the proportion of these three items marked by clerical markers was 19 per cent, 28 per cent and 48 per cent respectively. The way in which this data has been processed also has implications for supervisory data relating to these items (see 4.1.3.1).

Table 4.2 The mean (and median) number of seconds (across questions) taken to mark individual clips.

	Data Entry mean secs.	Clerical mean secs. (median)	Centre-based expert mean secs. (median).	Web-based expert mean secs. (median).
Paper 1	3.1	3.1 (3.0)	12.6 (11.0)	12.0 (10.2)
Paper 2	2.8	3.8 (3.5)	11.1 (10.4)	11.8 (10.5)
Mental A	2.0	2.3 (2.2)	-	-
Mental B	2.1	2.2 (2.3)	-	-
Mental C	2.0	2.0 (1.9)	-	-
Extension	-	-	41.2 (29.0)	37.5 (25.9)

Care should be taken in interpreting the mean (and median) figures, as, in some cases, large standard deviations betrayed considerable (within-question) variations in the time taken by markers to mark item clips. Unfortunately, data for data entry items was not available from the DWS Editor system at individual item level. However, NCS Pearson were able to supply data on the mean time taken by a marker to complete all the data entry items for one student, by paper and tier. The data entry figures in Table 4.2 have been derived by taking this figure and dividing by the number of items in the paper, aggregating the data for the four tiers within Papers 1 and 2.

Table 4.2 provides an 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 the timings were averaged across markers before being averaged across questions, thereby ignoring the substantial variance in the data and that different markers marked different numbers of clips for each question. However, it seems likely that the data still provide a reasonably useful approximation of the amount of time that it typically took to mark item clips. In calculating the means across questions, the times for conjoined clips have been averaged across the number of items within that conjoined clip.

The trends in Table 4.2 are reasonably clear. Across Papers 1 and 2, data entry and clerical markers took less time to mark item clips than expert markers, who tended to take around three to four times longer. This is understandable, as, in many cases, the expert questions would have required more attention and deliberation. The mean time per clip is calculated as the time elapsed from the appearance of an item on screen until the submission of a score, i.e., the 'thinking' time involved. It does not include the download time, which was generally considerably longer for home-based markers.[8] This explains why there was very little difference between centre-based and home-based markers. However, for Paper 1 and the Extension Paper the centre-based means were slightly longer. This may suggest that centre-based markers were taking advantage of the

[8] Following discussions with NCS Pearson, it is clear that the download time for home-based markers depended on a number of different factors. Some of these related to the markers' home computer equipment and their internet access, including processor speed and memory capacity of the markers' computers, type of connection to the internet, service provider and the time of day the internet was accessed. Other factors related to the type of item being marked, for example the size of the image file and the amount of thinking time required. When download time exceeded thinking time, the performance of the system governed the rate at which home-based markers were able to mark. A full performance study is beyond the scope of this report.

opportunity to consult with colleagues on some items, and thus increasing slightly the mean marking time per item. Mental arithmetic questions took the least time to mark, requiring only two seconds on average, and, as would be expected, responses to the Extension Paper took the longest time to mark. For comparison, the medians on the 2001 Year 7 Progress test were three seconds per clip on both papers for data entry and clerical markers and nine seconds (Maths A) and six seconds (Maths B) for expert markers.

Tables 4.2.1A to 4.2.2bA present marking speed data from the pilot, aggregated across markers 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. As can be seen from these three tables, on Paper 1 and 2, the longest mean times per clip for clerical markers were six seconds for Q17a/b and eight and seven seconds respectively for Q47a/b/c and Q57a/b/c. As these were conjoined clips, the times are not unexpected given that the markers were allocating a mark for each element. For single clips, the longest mean time recorded was five seconds for Q43d (Paper 2), a question for which students had to give four responses for two marks. For centre-based expert markers, the longest means were 29 seconds for both Q28b and Q54b/c/d and, for home-based experts, 27 seconds for Q30c and 31 seconds for Q54b/c/d. Both Q28b and Q30c were two mark items and Q54b/c/d was a three-mark item, requiring students to make an accurate drawing of a net. This last item was also one of three expert items where a decision was made midway through the marking period to allow clerical markers to mark them. It is possible that clerical markers took longer to mark this item than experts, thus contributing to a longer mean time for marking this item.

The 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 could be used to identify 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 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. A focus on speed also ignores the complexity of items being marked.

Unfortunately, the productivity data for clerical and expert markers was not split by tier. However, by multiplying the number of data entry, clerical and expert clips in each tier by

the mean time per clip for the paper as a whole (using the centre-based means for the expert markers), we can arrive at a very rough estimate of the notional average time that it took to mark each script on-line. These values are presented below in Table 4.3.

Table 4.3 A very rough estimate of the notional average time taken to mark each script on-line during the pilot.

	Data Entry	Clerical	Expert	Total
	mean secs. (no. clips)	mean secs. (no. clips)	mean secs. (no. clips)	Time (mins.)
Paper 1 – 3-5	3.1 (7)	3.1 (22)	12.6 (15)	4.6
Paper 1 – 4-6	2.8 (7)	3.1 (12)	12.6 (22)	5.6
Paper 1 – 5-7	3.5 (4)	3.1 (8)	12.6 (24)	5.7
Paper 1 – 6-8	3.6 (2)	3.1 (3)	12.6 (25)	5.5
Paper 2 – 3-5	2.4 (11)	3.8 (11)	11.1 (15)	3.9
Paper 2 – 4-6	3.0 (7)	3.8 (8)	11.1 (21)	4.7
Paper 2 – 5-7	2.8 (9)	3.8 (9)	11.1 (22)	5.1
Paper 2 – 6-8	3.1 (8)	3.8 (8)	11.1 (22)	5.0
Mental A	2.0 (25)	2.3 (5)	-	1.0
Mental B	2.1 (25)	2.2 (5)	-	1.1
Mental C	2.0 (26)	2.0 (4)	-	1.0
Extension	-	-	41.2 (12)	8.2

From these figures, it is possible to estimate that the time taken to e-mark all the scripts for one pupil was approximately 10-12 minutes depending on the tier and that the marking of expert items took approximately 6-9 minutes per pupil. However, it must be remembered that the figures in Table 4.3 do not include the time for the clip images to download, i.e., the delay between the submission of a score and the appearance of the next item on screen. The total marking time would therefore be somewhat higher than these figures. According to the *Questionnaire Survey of External Markers, 2002: Final Report*, September 2002, the mean time to mark a set of KS3 mathematics scripts conventionally was reported to be 15 minutes (excluding clerical tasks) and 20 minutes (including clerical tasks). The above estimates suggest that the total time taken for centre-based on-line marking of a complete set of scripts may be similar to that of conventional marking, once download time has been included. However, the advantage of the new technology system is that it enables a much better utilisation of scarce human resources, i.e. the expert markers.

Tables 4.1.1bA and 4.2.3A present data that encapsulate Cycle 2, the marking stage of the pilot. Table 4.2.3A shows the total number of scripts marked by paper and the number of items double marked by each category of marker.

4.1.3 Supervisory and adjudication demands

4.1.3.1 Supervision

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

In the 2002 pilot, markers could refrain from marking a particular response that they were unsure about, by sending it to a supervisor. Table 4.4 is a summary of data presented in 4.3.1A to 4.3.3bA (which capture the same information but for individual questions of each paper)[9]. Initially, in the original Table 4.3.2A submitted to NFER, there were virtually no clips at all reported as having been sent from clerical markers to supervisors without marking. However, this did not appear plausible, particularly in view of the NFER team's observation of the e-marking process that had taken place, where such activities had been seen and noted. An investigation at NCS Pearson revealed an error in the underlying process producing this data, leading to a revision of all the management and measurement data in respect of supervision and adjudication functions.

[9] In Tables 4.3.1A to 4.3.3bA, the number of clips referred to a supervisor is expressed as a percentage of the number of items processed. However, as the number of clips referred reflect individual clips, the percentage should have been expressed in relation to twice the number of items processed (since both markers could refer a clip for the same item). This adjustment has been made in Table 4.4.

Table 4.4 The median[10] (across questions) of the total number of clips sent to supervisors by markers.

	Data entry		Clerical		Centre-based expert		Web-based expert	
	median no. clips	% clips	median no. clips	% clips	median no. clips	% clips	Median no. clips	% clips
Paper 1	27.0	0.4	22.5	0.4	15.0	0.5	20.0	0.8
Paper 2	24.0	0.4	27.0	0.5	24.0	0.6	11.0	0.7
Mental A	18.0	0.3	24.0	0.4	-	-	-	-
Mental B	6.0	0.3	4.0	0.2	-	-	-	-
Mental C	23.5	0.7	8.0	0.2	-	-	-	-
Extension	-	-	-	-	3.0	8.0	3.0	11.5

In tables 4.3.1A to 4.3.3bA, the number of times a conjoined clip was referred to a supervisor is the same figure for each item within the clip. Unfortunately it was not possible to quantify whether the query was on one or more items on the same clip. For example, in one case the marker may have been unsure how to mark Q8a, whereas in another case s/he may have had a query on all three items. In both cases s/he would have had to send the conjoined clip Q8a/b/c. Although there were only a small number of conjoined clips, as these figures have not been adjusted, it is likely that the above figures (for Paper 1, Paper 2 and the Extension Paper) overestimate slightly the number of items sent to supervisors.

As can be seen from the above table, the number of clips sent to supervisors by data entry and clerical markers was generally below one per cent of the total clips viewed. On the mental arithmetic paper, Mental C, the percentage of clips referred to supervisors by data entry markers was noticeably higher on four items. One of these items (Q168) was identified in Study 6 as having a large number of responses BtCIA (26 per cent). Another item (Q145) with a high proportion of referred clips, required the correct response of 29,028, which (due to its length) may also have resulted in responses BtCIA. Given that students taking the Mental C Paper may also have poor handwriting, this is an issue to consider in any scale-up of the project.

[10] The median is shown in preference to the mean to avoid distortion, where the group of items is relatively small, e.g. mental arithmetic items marked by clerical markers.

Overall, home-based expert markers sent more items to supervisors than their centre-based counterparts. This is to be expected, in that the home-based markers had no opportunity for immediate personal discussions with colleagues or supervisors. On Papers 1 and 2, the items with the highest proportions of clips sent to supervisors by centre-based markers were for Q51a/b/c and Q69c (Paper 2)[11]. The former was a conjoined clip with a scale and three questions below it. The latter question was in tier 6-8 and required students to carry out a calculation, showing all their working. Both of these questions had a high incidence of responses BtCIA in one of the tiers in Study 6 (28 and 22 per cent respectively). For home-based markers, several items had relatively high numbers of clips sent to supervisors unmarked (Q26b, Q30c, Q36a and Q36b in Paper 1 and Q69c and Q72a in Paper 2). All of these items were from higher tier papers and, apart from one item, involved students showing their working. The other item was where students had to draw a response using a straight edge and compasses, and an overlay was provided to assist in the marking of this item.

Although only small numbers of extension papers were marked, and by senior markers only, a high proportion of student responses were referred to a supervisor, indicating that double-marking of these items could produce a high level of workload for supervisors in a scaled-up operation.

Table 4.3.4A records the number of messages that were sent by markers to supervisors, and by supervisors to markers, attached to or in response to referred clips[12]. Clerical markers sent approximately 350 messages relating to clips from Papers 1 and 2, but very few messages relating to mental arithmetic clips. Both centre-based and home-based expert markers sent a total of approximately 600 messages for each of Papers 1 and 2. The home-based supervisor sent approximately 190 messages to his marking team in

---

[11] As already reported, both the productivity and supervision figures for clerical markers and expert markers are incorrect in respect of three questions on Paper 2 (Q52a, Q54b/c/d and Q66a). The data for these items were included as if marked by experts, whereas the marking was split between expert and clerical markers. This has resulted in a strange anomaly in the number of clips referred to supervisor for these items in Tables 4.3.3A and 4.3.3b, in particular with regard to Q54b/c/d. It is assumed that the figures quoted on these items include the items transferred to clerical markers.

[12] The information provided in Table 4.3.4A in respect of the messages sent to and from markers was revised several times by NCS Pearson and some concerns about the accuracy of the data presented in this table remain. For data entry markers, there is no separate messaging function within the DWS Editor system so the data represents the total number of clips referred to supervisors.

response to messages accompanying items 'sent to review', i.e., referred to him by his markers. The home-based supervisor also communicated to his markers via group e-mails (although the number of these was not recorded), reminding them to read these using 'The Message of the Day' facility within e-PEN. In contrast, the centre-based supervisor sent no electronic messages at all to her team, preferring to talk to her markers face-to-face about marking issues. It is possible that the number of messages sent indicates differences in style amongst the supervisors or that, having discussed the home-based marking, which occurred first, it was decided that sending messages electronically was unmanageable. However, it may suggest a considerable difference in workload between supervisors operating at a marking centre and those supervising a home-based team.

4.1.3.2 Adjudication

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

In the 2002 pilot, two clips were marked, and therefore two marks were provided for each item. Hence, it was necessary for a supervisor 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 data entry, clerical, centre-based expert and home-based expert markers are presented in Tables 4.4.1A, to 4.4.3bA. The data are summarised in Table 4.5.

Table 4.5 The median (across questions) of the total number of items that required adjudication.

	Data entry		Clerical		Centre-based expert		Web-based expert	
	median no. items	median % items	median no. items	median % items	median no. items	median % items	median no. items	median % items
Paper 1	80	2.3	31	1.4	55	4.2	51	4.8
Paper 2	44	1.5	40	2.1	78	4.9	49	4.9
Mental A	65	2.2	25	0.9	-	-	-	-
Mental B	26	2.4	8	0.7	-	-	-	-
Mental C	48	2.9	12	0.7	-	-	-	-
Extension	-	-	-	-	6	30.0	4	30.8

Once again the presence of conjoined clips in Papers 1 and 2 and the Extension Paper may have resulted in some slight overestimates of the numbers of items requiring adjudication. In addition, the figures for all papers may have been inflated by the problem identified during the observation of the e-marking, that of confusion about the protocol for null

responses. Thus, where the one marker correctly submitted a score of 'BL' for a missing response ('\' for data entry items) and the other marker used a zero, the resultant pair of scores would have required adjudication, even though there was no mark difference.

For data entry markers, the median of the total number of items adjudicated ranged from 1.5 per cent (Paper 2) to 2.9 per cent (Mental C). At an individual item level, items requiring adjudication ranged from 0.05 per cent (Paper 2, Q60b) to 17.3 per cent (Paper 2, Q71c) of the total number of items processed. For clerical items, the medians ranged from 0.7 per cent to 2.1 per cent at paper level and from 0.2 per cent to 4.8 per cent at individual item level. The three individual clerical questions with the highest number of adjudications were all conjoined clips involving multiple marks.

The number of expert items requiring adjudication was approximately four to five per cent across Papers 1 and 2. However, there was considerable variation at individual item level. The number of items referred to adjudication ranged from one per cent (Paper 1, Q02b) to approximately 30 per cent of Q21c/d (Paper 1). Q21c/d required students to give reasons why a particular method might not give good data, i.e., it required an explanation rather than a more straightforward numerical response. It is likely that differences between markers might occur more frequently on such items. However, figures as high as 30 per cent are worrying, as they suggest a potentially ineffective use of human resources (i.e., marked by two markers *and* by a supervisor).

Although only a small number of extension papers were marked, a large proportion of the items on this paper required adjudication. The median across questions was 30 and 31 per cent of items for centre-based and home-based markers respectively. At an individual item level the number of item pairs in disagreement ranged from zero to 70 per cent of the total number of items processed. However, the adjudication figures for the Extension Paper are distorted by the fact that the majority of items on this paper were scanned as conjoined clips. Such clips had between two and six mark pairs, where possible disagreements could occur. Where any one mark pair was in disagreement, the adjudication was recorded against each item within the clip. For example, for Q172a/b/c/d/e/f, the number of items referred to adjudication was recorded as 14 of the 20 items processed, i.e. 70 per cent. However, the true number of disagreements could have been anywhere between 12 per cent (14 of 120 possible mark pairs in disagreement) to 70 per cent (84 of 120).

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. For further discussion of this issue, see 5.1.4.

Adjudication could also be performed by 'back-reading' the marking of a particular item or marker and changing the mark(s) awarded as necessary. Tables 4.5.1aA and 4.5.1bA show the number and percentage of items that were changed by back-reading. These figures are extremely small and, unfortunately, there is no comparison information as to how many items were viewed using the back-read function and remained unchanged. Anecdotally, both the Chief e-Marker and the two Senior e-Markers said during the observation period that they had spent very little time back-reading, due to the need to prioritise the review and adjudication functions.

4.1.4 Processing accuracy

Issue logs in respect of processing errors were produced by NCS Pearson for Cycle 1, from the receipt of scripts from schools to their despatch to conventional markers. Generally, scripts appeared to be processed effectively, and where errors did occur they were quickly identified and dealt with. The location of scripts within the system was facilitated by the logging of script barcodes at all key stages of the process.

4.2 Analysis of the measurement data

The focus of the analysis of measurement data was the accuracy of the results arising from the electronic marking. The data collected during the pilot allowed the calculation of the following:

- paper and subject-level correlation between conventional marks and e-marks;
- paper and subject-level absolute mark differences between conventional marks and e-marks;
- question-level concordance statistics;
- subject-level concordance statistics.

The data underlying the following analyses are presented in Appendix 4.2.

4.2.1 Methodology

As described in 2.2.3, scanning of student scripts was carried out in advance of the conventional marking. Therefore, as marks awarded to responses were not visible on item clips, the on-line marks were awarded independently of the conventional ones. In order to provide data for comparative purposes, the marks that had been awarded by the

conventional markers were input by the PECS data capture bureau.[13] This enabled a valid comparison of marking consistency between the conventional and the new technology systems, i.e. between system comparisons.

As a double e-marking model was employed in the 2002 pilot, this made it possible to consider the consistency of marking within the new technology system, i.e. between-marker comparisons for data entry, clerical and expert items. As in 2001, the 2002 evaluation examines the kinds of measurement data that the NCS Pearson new technology system can routinely yield and provides further baseline information on the quality of on-line marking to inform future pilots and trials.

4.2.2 Inferential caveats

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

4.2.2.1 The nature of the data

All of the analyses presented below are based upon data that were provided for the NFER, by NCS Pearson, as specified in the *Framework for Evaluating the New Technologies Pilot 2002*. The following information was requested:

1. eleven marker-level databases, containing item-level between-marker reliability data – the number of clips for which the marker agreed with a re-marker;

2. eleven marker-level databases, containing item-level absolute-marker reliability data – the number of clips for which the marker agreed with the Chief e-Marker;

3. twenty-four student-level databases, corresponding to each of the 12 papers, containing item-level mark data from both the conventional and the on-line process;

4. student-level and marker-level databases, corresponding to the four papers marked conventionally and electronically for Study 5.

---

[13] As PECS 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.

4.2.2.1.1 The marker-level databases (between-marker reliability)

Initial checking of the marker-level databases uncovered some errors and anomalies throughout.

- The item quantities in columns 'A' (agree), 'D' (disagree) and 'O' (no mark awarded) did not equal the total number of clips viewed by the marker.

- There were errors/omissions in the 'O' column, i.e. items sent to a supervisor without marking. (The information in this column represented the number of items for which the marker sent the clip directly to a supervisor without marking and had been requested in order to contextualise the basic 'agree versus disagree' data. As reported earlier, this error also affected the management data in Tables 4.3.1A to 4.3.3bA.)

- Marker IDs were not unique and some markers had more than one ID. (This had occurred in the 2001 pilot and was specifically mentioned in the framework document as something that was to be avoided in 2002. Clear instructions in the detailed data specifications were that marker IDs must be consistent across data bases and that no marker should be given more than one ID number.)

Following discussions with NCS Pearson, the data were revised and resubmitted to NFER. However, the number of errors unfortunately brings into question the level of confidence with which the final data can be interpreted. Due to pressure of time, the issue of marker ID numbers was resolved by NCS Pearson submitting a list of names relating to the various marker IDs so that the marker data could be aggregated by NFER. (Supervisor data was left at a dis-aggregated level due to problems of identification.)

As already reported (see 4.1.2), the reliability data for clerical markers and expert markers are incorrect in respect of three questions on Paper 2 (Q52a, Q54b/c/d and Q66a). The data for these items were included as if they had been marked by experts, whereas the marking was split between expert and clerical markers. Further outstanding technical caveats need to be taken into account in respect of the reliability data. Firstly, a bug in the e-PEN system resulted in 2319 items being only single-scored. Thus, for those items, the figures in the reliability data will not equate to twice the number of items. Secondly small numbers of validity items that were introduced but not escalated appear in the data as scored items alongside genuine student responses. Finally, for multitrait items (conjoined clips) each trait will have the same number of send to review counts, as it is impossible to distinguish which trait was referred.

4.2.2.1.2 The marker-level databases (absolute-marker reliability)

Initially, as with the between-marker data, there were errors/omissions in the 'O' column, i.e. items sent to a supervisor without marking, in the validity databases. NCS Pearson corrected these errors. However, a further issue concerned the number of databases, specified within the framework document, for which no data was produced. No validity data was provided for clerical markers in respect of Paper 1 and the three mental mathematics papers, or for expert markers, in respect of the Extension Paper.

NCS Pearson reported that, as the Extension Paper was to be marked by senior markers or above, the Chief and Senior e-Markers had felt that there was no necessity for validity items. With regard to clerical markers, NCS Pearson stated that there was no requirement for validity data in the initial customer requirements document. However, as tables for these databases were clearly included in the framework document produced in May 2002 (Tables 30.1 and 32.1 to 35.2b), clarification should have been sought on this issue. As validity items were included (and data subsequently produced) for some clerical items (within Paper 2), it is not clear whether there was an intention to produce validity data for clerical markers or not. In the live e-marking period, there were problems loading validity items into the system (see section 2.2.7), which may explain the lack of data, particularly in respect of Paper 1. Even where validity data has been supplied, it is of fairly limited quantity in many cases. Thus there may be some uncertainty in the interpretation of this data.

4.2.2.1.3 The nature of the sample

Table 4.6 shows the representation of the pilot sample at school level.

Table 4.6 Sample representation (schools)

		population		sample	
		Number	%	Number	%
LEA type	London Borough	703	13	7	9
	Metropolitan Authorities	1096	21	21	27
	English Unitary Authorities	864	16	10	13
	Counties	2615	50	39	51
Region	North	1449	27	28	36
	Midlands	1633	31	21	27
	South	2196	42	28	36
Size of age group	1-80	2055	39	2	3
	81-160	1223	23	26	34
	161-240	1486	28	36	47
	241+	514	10	13	17
Total schools		5278	100	77	100
Achievement band (KS3 Maths performance)	Lowest band	1059	29	23	30
	2nd lowest band	650	18	17	22
	Middle band	622	17	16	21
	2nd highest band	601	16	9	12
	Highest band	747	20	12	16
Total schools		3679	100	77	100

Since percentages are rounded to the nearest integer, they may not always sum to 100.

Chi-square tests were carried out, which showed that the only statistically significant difference between the characteristics of the pilot sample and the national population was in respect of the 'size of the age group' variable. Thus, schools with age groups of 80 pupils or fewer were under-represented in the sample. However, in terms of LEA type, region and bands of achievement in KS3 mathematics, the pilot sample was appropriately reflective of the national picture.

4.2.2.2 The validity of the data

Having clarified the nature of the data and sample that provided the basis for Study 4, 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.

4.2.2.2.1 The validity of conventional marks

Unlike last year, scripts in the 2002 pilot had been checked and borderlined before marks were input by PECS for NCS Pearson. As such checks had been undertaken, the conventional data should represent 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.

4.2.2.2.2 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. In the 2001 pilot the markers were far less experienced than the typical national curriculum marker. By contrast, in 2002 the markers used for on-line marking were far more experienced than the typical national curriculum marker. This is again 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.

As last year, there was again a problem with elements of students' responses to questions occasionally being recorded beyond the clip image areas (see Section 2.3). These elements were simply not available to e-markers for scrutiny. To the extent that students' 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.

4.2.3 Between-system comparisons

The principal comparison of results between conventional marking and on-line marking concerned final marks. Agreement between mark totals from the two systems was

investigated using correlation and mean mark difference statistics.[14] The extent of agreement between marks awarded for each item of each script was subsequently analysed using concordance statistics.

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. This is because there was no independent external reference point determining the 'correct' mark for each script or item.

4.2.3.1 Correlation

Coefficients of correlation essentially indicate the degree of consistency between two sets of data; in this case the marks awarded to individual students during the conventional and new technology processes, respectively. These were considered for each of the papers separately. The results of these analyses are presented in Table 4.7.[15]

[14] 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).

[15] The correlation for Mental Arithmetic B is inaccurate due to an error in the new technology data that was reported to the NFER on 12 November 2002.

Table 4.7 The correlation between conventional and new technology marks

	tier		centre-based	web-based	overall
Paper 1	3 to 5	coeff	0.9953	0.9902	0.9934
		sig	0.000	0.000	0.000
		N	975	652	1627
	4 to 6	coeff	0.9899	0.9888	0.9894
		sig	0.000	0.000	0.000
		N	1008	670	1678
	5 to 7	coeff	0.9763	0.9827	0.9794
		sig	0.000	0.000	0.000
		N	861	576	1437
	6 to 8	coeff	0.9847	0.9900	0.9870
		sig	0.000	0.000	0.000
		N	368	252	620
Paper 2	3 to 5	coeff	0.9855	0.9614	0.9756
		sig	0.000	0.000	0.000
		N	953	646	1599
	4 to 6	coeff	0.9907	0.9877	0.9895
		sig	0.000	0.000	0.000
		N	968	648	1616
	5 to 7	coeff	0.9800	0.9873	0.9834
		sig	0.000	0.000	0.000
		N	802	552	1354
	6 to 8	coeff	0.9867	0.9899	0.9882
		sig	0.000	0.000	0.000
		N	350	227	577
Mental	A	coeff	0.9898	0.9944	0.9919
		sig	0.000	0.000	0.000
		N	1594	1052	2646
	B	coeff	0.9740	0.9959	0.9826
		sig	0.000	0.000	0.000
		N	624	418	1042
	C	coeff	0.9954	0.9853	0.9915
		sig	0.000	0.000	0.000
		N	965	635	1600
Extension		coeff	0.9653	0.9623	0.9627
		sig	0.000	0.000	0.000
		N	20	13	33

Most of the reliability coefficients are 0.98 or higher indicating a high degree of agreement between the conventional marks and the new technology marks. On Papers 1 and 2 the only coefficient below this level was the correlation between the marks of students in the home-based pool for Paper 2, Tier 3-5. The standard of reliability for the extension paper was slightly lower than for the main papers at 0.96.

Tables 4.6.1A to 4.6.3A give the correlation coefficients between conventional and new technology marks at overall student level, i.e. for all the combinations of papers taken by students in the pilot. All of these correlation coefficients were 0.98 or higher. The findings are therefore very similar to those of the 2001 pilot, in which the correlation

coefficients of the three individual maths papers were 0.98 and the overall subject level correlation was 0.99. The high correlation coefficients seem to provide evidence that mutually validates both systems of marking for mathematics.

4.2.3.2 Mean mark difference

Although the correlation between conventional and new technology marking was very high, it is conceivable that the task demands of the two systems 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 were also investigated. The possibility of general lenience or harshness was evaluated at paper level and student level (i.e. by analysing the mean mark difference between conventional and new technology marks of students taking the nine different combinations of papers). The full data are presented in Tables 4.7.1A to 4.7.4A and the student level data are summarised in Table 4.8 .[16]

Table 4.8 The mean mark difference between conventional and new technology marks – (CM minus NT).

Combination	Paper 1	Paper 2	Mental	Extension	Mean mark difference	Number
1	3 to 5	3 to 5	C	N/A	0.48	1365
2	4 to 6	4 to 6	A	N/A	0.65	949
3	4 to 6	4 to 6	B	N/A	0.87	356
4	5 to 7	5 to 7	A	N/A	0.85	738
5	5 to 7	5 to 7	B	N/A	1.32	331
6	6 to 8	6 to 8	A	N/A	0.76	327
7	6 to 8	6 to 8	B	N/A	1.53	106
8	6 to 8	6 to 8	A	Extension	0.16	25
9	6 to 8	6 to 8	B	Extension	-	-

Although the mean mark differences were generally small, across the student pool as a whole, the mean total scores for conventional marking were always higher than the new technology marks (i.e. when new technology marks were subtracted from conventional marks the difference was positive). However, in contrast to the 2001 pilot, none of the mean mark differences revealed by this analysis were statistically significant ($p < 0.05$) either at student level, as shown above, or at individual paper level (see Tables 4.7.1A to 4.7.4A).

[16] The mean mark differences for combinations involving Mental Arithmetic B are inaccurate due to an error in the new technology data that was reported to the NFER on 12 November 2002.

Although the differences were not statistically significant, it is interesting to note that the direction of the difference, in favour of conventional marks is consistent, replicating the finding from the 2001 pilot. As reported last year, this suggests that the process of new technology marking may result in the awarding of marginally lower marks than the process of conventional marking. The explanation for this finding is not clear-cut and could be attributed to a number of factors. One possibility might be that conventional markers – seeing personalised scripts rather than de-personalised question responses – were sub-consciously more disposed to give the benefit of the doubt to candidates. Another possibility is that differences in the calibre of the e-marking team compared to the general marking population may have may have resulted in the latter group being more lenient than the former. A further possibility is that the items that could not be marked electronically, because part of the student's response was beyond the clip image area, may have resulted in the award of additional marks when those same responses were marked conventionally.[17] More evidence to support or refute this possibility may be available from NCS Pearson, who have been evaluating how many of the BtCIA responses identified in Study 6 would have impacted on the overall mark. However, as reported above, the resultant differences were not statistically significant and would not therefore pose a significant threat to a scale-up of the new technology system for mathematics tests.

4.2.3.3 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 4.9 summarises data presented in Tables 4.8.1A to 4.8.8A. 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 item than on a single mark item.

---

[17] BtCIA responses were coded as missing values in calculating the mean scores for new technology marking. However, these items were not excluded from the comparison between conventional and new technology marking.

Table 4.9 Mean percentage exact agreement figures (the number of questions from which each average was computed is in brackets).

	Data entry	Clerical		Expert			
	1	1	2	1	2	3	4
Paper 1							
3 – 5	99.5 (7)	99.2 (19)	98.0 (4)	97.2 (16)	97.6 (4)	88.8 (1)	-
4 – 6	99.4 (7)	99.3 (11)	98.0 (2)	95.2 (25)	97.4 (7)	-	-
5 – 7	99.3 (4)	99.3 (8)	98.6 (1)	93.2 (20)	96.4 (11)	91.9 (1)	-
6 - 8	99.7 (2)	98.2 (2)	98.2 (1)	93.1 (18)	94.1 (13)	86.9 (2)	93.1 (1)
Paper 2							
3 – 5	99.1 (11)	97.7 (11)	97.9 (4)	95.4 (12)	97.1 (7)	92.2 (1)	-
4 – 6	99.5 (7)	99.1 (10)	99.0 (2)	95.6 (17)	96.9 (7)	94.7 (1)	91.8 (1)
5 – 7	99.6 (9)	98.8 (12)	-	95.2 17	95.0 (7)	89.0 (2)	89.7 (1)
6 - 8	99.6 (8)	99.3 (11)	-	94.4 (13)	95.5 (6)	86.5 (4)	91.9 (1)
Mental A	99.4 (25)	99.2 (5)	-	-	-	-	-
Mental B	98.1 (25)	99.2 (5)	-	-	-	-	-
Mental C	98.9 (26)	98.8 (4)	-	-	-	-	-
Extension	-	-	-	93.9 (15)	92.3 (11)	90.9 (1)	-

These data complement and extend the correlation coefficients presented earlier. Once again, the figures for these mathematics questions were generally high. The average percentage agreement across the clerical and data entry questions did not fall below 98 per cent[18]. For the expert questions, the percentage agreement ranged from 87 per cent for three-mark items in Tier 6-8 to 98 per cent for two-mark items in Paper 1, Tier 3-5. Generally the level of exact agreement tended to be lower in the higher tier papers

It is likely that the item-level information in Tables 4.8.1A to 4.8.12A will be of particular interest to senior markers and test development teams. By examining the characteristics of questions with poor concordance it may be possible to identify patterns of marking

[18] The percentage exact agreement figure for data entry items in Mental Arithmetic B is inaccurate due to an error in the new technology data that was reported to the NFER on 12 November 2002.

error that can feed into future developmental work. It is interesting to note that some of the items showing poor concordance between conventional and new technology marks were items identified as having a high frequency of BtCIA responses in Study 6 and this is therefore likely to be one of the factors contributing to marker disagreement on individual items.

4.2.3.4 Subject-level concordance

Finally, to provide an overall comparison of conventional and new technology marking, the absolute agreement between the levels arising from the two approaches was examined. Table 4.10 compares the percentage of students at each level using the conventional and new technology marks.

Table 4.10 The difference between levels awarded using conventional and new technology marks

	Conventional marks	New technology marks
Percentage of students at level N	1.5	1.6
Percentage of students at level 2	1.2	1.2
Percentage of students at level 3	12.4	12.4
Percentage of students at level 4	21.5	21.9
Percentage of students at level 5	24.9	24.2
Percentage of students at level 6	23.2	22.5
Percentage of students at level 7	12.3	13.1
Percentage of students at level 8	3.0	3.1
Percentage of pupils with different levels when using NT instead of CM	4.9	

n = 4172

Table 4.10 shows that five per cent of students would have received different levels had the new technology marks been used rather than the conventional ones. Four per cent of students would have received a lower level, while one per cent would have received a higher level.

4.2.4 Within-system comparisons

In addition to analyses comparing the quality of marking between systems, due to the double-marking model adopted in the 2002 pilot, it was also possible to conduct analyses to compare the quality of clerical and expert marking within the new technology system (i.e., between the on-line markers).

Three sets of databases were collected for this purpose:

1. between-marker reliability databases for clerical markers;

2. between-marker reliability databases for home-based expert markers

3. between-marker reliability databases for centre-based expert markers.

As well as collecting data on the percentage of agreement between markers, data was also collected on the number of items passed to supervisors without marking. It should be noted that although the basic agree/disagree statistics provide important information, it is more valuable when contextualised by this second set of data. If a marker considers only those items for which s/he is certain of the mark, s/he 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/his marking statistics would not be directly comparable with those from another, less conservative, marker.

4.2.4.1 Between-marker reliability

Tables 4.9.1A to 4.9.4A (Appendix 4.2) present data concerning the performance of individual markers. For each paper, these include the total number of clips marked across all questions, the percentage of agreement and disagreement between each marker and the various re-markers and the percentage of clips for which the marker did not award a mark[19]. These data provide some indication of the difference in marking quality between the markers. The between-marker reliability data are summarised in Table 4.11.

[19] 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. Although care was taken to assign expert markers to groups of questions with similar marking characteristics, it is possible that some items were harder to mark consistently than others. Therefore any comparisons between markers should be made cautiously. This is another issue that future contractors will have to consider when developing marker reliability statistics for monitoring purposes.

Table 4.11 Between-marker reliability data.

	Clerical		Centre-based expert		(Home) Web-based expert	
	Percentage exact agreement	Percentage sent to supervisor	Percentage exact agreement	Percentage sent to supervisor	Percentage exact agreement	Percentage sent to supervisor
Paper 1	98.2	0.6	94.8	0.8	94.1	1.1
Paper 2	98.1	0.6	94.1	1.5	92.6	2.2
Mental A	98.7	0.5	-	-	-	-
Mental B	98.8	0.4	-	-	-	-
Mental C	99.0	0.4	-	-	-	-
Extension	-	-	81.2	4.3	84.2	8.2

The percentage agreement figures between clerical markers were consistently high. As can be seen in Table 4.11, the overall level of agreement between clerical markers was 98 per cent for Papers 1 and 2 and 99 per cent for the three mental arithmetic papers. The lowest percentage agreement for any individual clerical marker was 97.5 per cent (see Table 4.9.1A). The percentage of items, which were sent to supervisors by clerical markers with no marks awarded, was less than one per cent on each of the five papers.

For centre-based expert markers, the level of exact agreement was 95 and 94 per cent respectively on Papers 1 and 2, and 81 per cent on the Extension Paper. The overall agreement figures for home-based expert markers were slightly lower for Papers 1 and 2 but higher on the Extension Paper. As reported earlier, on all three papers, home-based experts sent more unmarked items to supervisors than centre-based experts. As can be seen from Tables 4.9.1A to 4.9.4A, there were considerable differences between individual expert markers, both in the level of agreement with other markers and in the numbers of items sent to a supervisor unmarked. For example, on Paper 1, the level of agreement achieved by individual markers ranged from 86 per cent (home marker 161) to 98 per cent (centre marker 124). On the same paper, the level of disagreement with re-markers ranged from two to 13 per cent and the percentage of items sent to supervisors unmarked ranged from less than one per cent to six per cent. Generally, the supervisors had a much higher level of disagreement with re-markers. This is because they were only marking unmarked clips sent by one of a pair of markers, i.e. those that have been found difficult to mark for whatever reason. It is therefore to be expected that these would be more likely to cause disagreement. The overall percentage disagreement figures are therefore similar to the figures in Table 4.5 – the percentages of items requiring adjudication. The level of error that could potentially be removed if a double-marking model were adopted in a new technology system would be approximately half of the

disagreement figure, (assuming that in a single marking system, each marker would probably be correct in 50 per cent of cases).

Between-marker reliability data facilitates comparisons between markers. For example, on Paper 2 one centre-based marker (ID 125) marked approximately 14,000 clips. S/he agreed with the re-markers on 90 per cent of those clips and sent 5 per cent of her/his clips to a supervisor unmarked. Contrast this with marker 124, who marked over 17,000 clips in total, achieved a higher agreement level of 95 per cent, and sent less than two per cent of his/her clips for review. The availability of data in this level of detail can be informative but obviously would need to be interpreted with caution. A further couple of examples may serve to illustrate this. Home based marker 144 achieved one of the lowest agreement levels on Paper 1 (88 per cent). However, on Paper 2 s/he achieved an agreement level of 97 per cent – one of the highest for that paper. S/he sent over seven per cent of items for review on Paper 1 but less than two per cent on Paper 2. Centre-based marker 117 achieved percentages of exact agreement with re-markers of 93 and 94 per cent for Papers 1 and 2 respectively, in each case with very low percentages of items sent to supervisors unmarked. Which marker is the more reliable? This question may only be resolved when information at paper level is supplemented by an examination of the reliability data on individual items. This could help to identify whether a marker had been marking particularly difficult items, relative to other markers, had a specific problem with one marking issue or had a poor level of agreement with other markers generally across all items. The type of information produced in this pilot could be particularly useful for comparing members of marking teams marking the same items (in similar proportions).

The availability of marker monitoring data 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. This pilot provides extensive baseline data to inform future decisions as to what levels of reliability would be acceptable for different items/papers. A further issue is to the extent to which monitoring data is used during the marking period itself, or whether it is reviewed after marking has finished.

4.2.4.2 Absolute marking standards (validity)

Each marker was compared with the absolute standard determined by the Chief e-Marker through the use of 'seed' or validity items. As described in section 2.2.4, each expert marker received one validity clip in every 40 clips marked. The marker's response to each validity item was compared with the correct mark, as determined by the Chief e-

Marker. Thus absolute-marker reliability data sets were produced along the same lines as those produced for the between-marker reliability analyses.

Tables 4.10.1A and 4.10.2A give the validity data concerning the performance of individual markers. For each marker, these tables show the total number of clips marked/viewed for validity purposes, the percentage of agreement and disagreement between the marker and the Chief e-Marker, the percentage of clips for which the marker sent the item to a supervisor without awarding a mark and the total number of clips marked/viewed for this question (excluding validity clips)[20]. These data are summarised in Table 4.12 below.

Table 4.12 Absolute-marker reliability data[21].

	Clerical		Centre-based expert		(Home) Web-based expert	
	Percentage exact agreement	**Percentage sent to supervisor**	**Percentage exact agreement**	**Percentage sent to supervisor**	**Percentage exact agreement**	**Percentage sent to supervisor**
Paper 1	n/a	n/a	92.6	1.9	89.3	6.0
Paper 2	97.3	0.3	88.8	0.7	78.3	15.3
Mental A	n/a	n/a	-	-	-	-
Mental B	n/a	n/a	-	-	-	-
Mental C	n/a	n/a	-	-	-	-
Extension	-	-	n/a	n/a	n/a	n/a

Where the data are available, the absolute-marker reliability figures are lower than the corresponding figures for between-marker reliability. One possible reason for this is that the Chief e-Marker was picking particularly challenging student responses for validity items (i.e. items that may be more difficult to mark) rather than sampling responses from across the full range of student responses. This would be a perfectly legitimate tactic in order to assess, on a question by question basis, whether markers had understood some of

[20] The relationship between the number of clips marked/viewed and the number of validity clips should be in the approximate ratio of 40:1. As validity clips were not loaded for all items this relationship is not maintained once the data for each marker is totalled across questions. However, inexplicably some markers seem to have received more validity clips than would be expected, for example home-based marker 137 (Paper 2).

[21] As reported in section 4.2.2.1.2 no validity data was produced for the three mental arithmetic papers, the Extension Paper and Paper 1 (clerical markers).

the finer distinctions of the mark scheme. However, caution would have to exercised if such data was then totalled across groups of questions in order to compare markers, as not all questions might have the potential for setting 'difficult' validity items.

With regard to expert markers, the percentages of exact agreement are much lower for home-based markers than for centre-based markers, with higher percentages of validity items being sent to the home-based supervisor unmarked. Although this was also the case with the between-marker reliability data, the differences were not as marked. One possibility is that the quality of the home-based markers was of a lower standard than the centre-based team, in that they were much less confident in marking the more difficult validity items. This could be because they didn't have the face-to-face support of their supervisor and the Chief e-Marker. However, although the pre-loaded validity items were common to each pool, the majority of validity clips were escalated from within each database separately. It is therefore possible that the validity items in the home-based pool may have been more difficult than those in the centre-based pool. The implication for any scaled-up operation is that if absolute-marker reliability data is to be used to compare markers, this should be done at question level or at team level across markers who have marked the same questions, and therefore the same validity items. It is also recommended that validity items are carefully selected and pre-loaded so that there are sufficient appropriate items for validity purposes at the start of the marking period.

As shown in Tables 4.10.1A and 4.10.2A, the percentage level of agreement between individual expert markers and the Chief e-Marker ranged from 58 per cent to 98 per cent (excluding markers who marked less than 100 validity clips). In the former case, almost over 35 per cent of the validity clips were sent to the supervisor unmarked (Paper 2, home-based marker 139). Several markers had levels of exact agreement of less than 80 per cent; levels that may indicate a cause of concern for mathematics papers. (However, this may depend on whether these markers were marking particularly challenging or complicated validity items, compared to other markers.) Unfortunately if the validity rate is set at only one in every 40 items, there is insufficient information during the marking phase itself to quickly identify markers who are not marking to the required standard. In a double marking model, it is likely that most errors would be picked up by adjudication. However, if anything less than a double-marking model were to be adopted in a scaled-up operation, the validity function would be of much greater significance and a much higher rate would need to be considered.

4.2.5 Script annotation error

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. Mark totals from the front of the scripts were input and compared with the totals computed from the sum of the individual marks. This gave some indication of the level of errors caused by mistakes of addition by conventional markers. These marks would be subject to review, after checking by teachers.

Table 4.11A records the prevalence of script annotation errors by paper. They were least common for the Mental Arithmetic Papers (0.3, 0.9 and 1.5 per cent respectively) and for the Extension Paper, where none of the scripts had addition errors. Across the different tiers of Papers 1 and 2, the level of errors was between three and four per cent. For high-stakes tests, these figures are very high, particularly in view of the fact that clerical checks had already been carried out on these scripts prior to their return to schools. The level of errors was slightly less for the Y7 mathematics progress tests in the 2001 pilot (between two and three per cent), although there were fewer marks to total compared to the KS3 papers.

As suggested in the 2001 pilot report, it is possible that the frequency of script addition error is significantly under-represented in the national curriculum review data (i.e., that mark addition errors frequently go unreported – particularly when they favour, rather than penalise, pupils). Errors such as these would simply not occur at all during new technology marking. This represents a very positive argument in favour of e-marking.

4.3 Study 5 – a controlled evaluation of marking reliability

As part of the 2002 KS3 pilot, it was proposed that a controlled evaluation of marking reliability would be carried out using an experimental manipulation.

4.3.1 Methodology

Four of the centre-based e-markers were contracted to mark a set of 100 scripts (in addition to their basic conventional and electronic allocation). The four markers marked each of these scripts both conventionally and electronically. The specification for each marking stage was as follows.

4.3.1.1 Conventional marking

During Cycle 1, NCS Pearson were responsible for reproducing full-page images for:

- a random sample of 25 Paper 1 scripts from tier 3 to 5;
- a random sample of 25 Paper 2 scripts from tier 4 to 6;
- a random sample of 25 Paper 1 scripts from tier 5 to 7;
- a random sample of 25 Paper 2 scripts from tier 6 to 8.

Twenty-three schools were represented within the sample of 100 scripts and the sampling resulted in a good spread of scores within each tier.

The pupil scripts were reproduced as four identical sets of 100 scripts (in paper booklet form). The four centre-based markers each marked a set of these 100 scripts, soon after they had marked their initial allocation of live conventional scripts. However, as the marks from this study were not being returned to schools, the marks were neither checked nor borderlined. The data arising from this conventional marking was input by PECS, the NCS Pearson data entry supplier, as a separate exercise after the input of live data from the main pilot.

4.3.1.2 E-marking

NCS Pearson constructed a separate e-marking database into which a sub-set of items was loaded, corresponding exactly to the 100 scripts sampled for the conventional marking exercise. Responses to all items for all 100 students were included (not simply the expert items). Likewise, each of the four e-markers was required to provide marks electronically for all items for all 100 students. Note that this database used the regular item image clips and not full page images.

The e-marking for Study 5 followed immediately after the main e-marking phase. The clerical and expert items were mixed together and were presented to the markers in paper order. The e-markers marked in a manner analogous to the regular e-marking, that is, they marked all students' responses for the first question, then progressed to the second question, and so on until all responses to all questions for all 100 students had been marked (by each e-marker). Student responses for the data entry items were entered using DWS Editor, for which the expert markers received additional system training. There was no conventional refresher training for the clerical items or for expert items that they had not marked during the main e-marking period. Markers could not send items to review, i.e. leave an item unmarked, and therefore in cases where responses extended beyond the clip image area, they were asked to mark as seen. There was no adjudication for the additional exercise, in order that the statistics outlined below could be produced.

One caveat should be noted. The conventional marking (of the 100 scripts) took place before the main e-marking phase, whereas the e-marking for Study 5 was conducted after it. As such, increased reliability in the new technology marking might be, to some extent, a function of greater practice.

4.3.2 Analysis

This study enabled production of the following statistics:

1. between-marker reliability data for conventional marking;
2. between-marker reliability data for new technology marking;
3. within-marker reliability data for conventional versus new technology marking.

Neither analyses 1 nor 3 had been produced previously for key stage 3 mathematics, nor had analysis 2 been produced under such controlled circumstances.

4.3.2.1 Analysis of between-marker reliability for conventional and new technology marking

Between-marker reliability was investigated by means of comparisons of total marks and also by the level of exact agreement between the marks awarded for each individual item. However, these analyses do not indicate whether one marker was more or less accurate than the others were - only the extent of the agreement. This is because there was no independent determination of the 'correct' mark for each item.

4.3.2.1.1 Correlation

Coefficients of correlation were computed to show the degree of consistency between the four markers for the same 25 scripts. These were considered for each paper separately, for both the conventionally marked scripts and the e-marked scripts. Tables 4.13 and 4.14 present the correlation statistics for the two marking processes, by paper.

Table 4.13 The correlation between markers – conventional marking

	Marker	1	2	3	4
Paper 1, Tier 3 to 5	1	1.0000	0.9953	0.9934	0.9947
	2		1.0000	0.9948	0.9953
	3			1.0000	0.9977
	4				1.0000
Paper 2, Tier 4 to 6	1	1.0000	0.9814	0.9916	0.9875
	2		1.0000	0.9803	0.9859
	3			1.0000	0.9879
	4				1.0000
Paper 1, Tier 5 to 7	1	1.0000	0.9753	0.9838	0.9800
	2		1.0000	0.9770	0.9780
	3			1.0000	0.9810
	4				1.0000
Paper 2, Tier 6 to 8	1	1.0000	0.9871	0.9910	0.9905
	2		1.0000	0.9751	0.9835
	3			1.0000	0.9877
	4				1.0000

All based on N=25 and with p<0.00.

Table 4.14 The correlation between markers – new technology marking

	Marker	1	2	3	4
Paper 1, Tier 3 to 5	1	1.0000	0.9985	0.9976	0.9954
	2		1.0000	0.9973	0.9954
	3			1.0000	0.9968
	4				1.0000
Paper 2, Tier 4 to 6	1	1.0000	0.9819	0.9846	0.9791
	2		1.0000	0.9928	0.9815
	3			1.0000	0.9793
	4				1.0000
Paper 1, Tier 5 to 7	1	1.0000	0.9902	0.9870	0.9871
	2		1.0000	0.9880	0.9853
	3			1.0000	0.9832
	4				1.0000
Paper 2, Tier 6 to 8	1	1.0000	0.9878	0.9928	0.9869
	2		1.0000	0.9913	0.9870
	3			1.0000	0.9865
	4				1.0000

All based on N=25 and with p<0.00.

As might be expected when using a group of very experienced markers, the correlation coefficients for all the papers, marked both conventionally and using new technology were extremely high.

4.3.2.1.2 Exact agreement analysis

Marker reliability was further explored by examining the extent of exact agreement across each item marked conventionally, and then examining the level of exact agreement for the same items marked electronically. Tables 5.1.1A to 5.2.4A (Appendix 5) present the mean level of exact agreement across markers for each item and the percentage of items for which each marker agreed exactly with the three re-marking markers[22]. This information provides an indication of the difference in marking quality between the markers. In this particular study, the data showed that there was very little difference between the four markers when the marking of each one was compared with the other three. Averaging across items for each paper, the lowest average level of agreement between one marker and the rest was 94 per cent.

The mean percentage agreement data is summarised in Table 4.15.

Table 4.15 The mean percentage exact agreement (across markers) for conventional and new technology marking.

	Conventional marking	New technology marking
Paper 1, Tier 3 to 5	98.0	98.2
Paper 2, Tier 4 to 6	95.0	97.0
Paper 1, Tier 5 to 7	95.1	96.1
Paper 2, Tier 6 to 8	95.8	95.3

As can be seen from the above table, the average level of exact agreement between markers ranged from 95 to 98 per cent in both conventional and new technology marking. Comparing the findings from conventional marking and e-marking, there was very little difference in the level of agreement achieved, although in three of the four papers the level of exact agreement was slightly higher with the new technology marking. Similarly, in respect of individual items, the data suggests that levels of exact agreement were more dependent on the item itself than on the marking process. Thus, items marked conventionally with low levels of exact agreement (e.g. 73% for Paper 2, 4-6, Q48b) also tended to have low levels of exact agreement when marked with new technology (81%), and vice versa. An examination of items with less than 80 per cent exact agreement revealed that all of these items were either ones where students had to give an explanation

[22] The exact agreement was computed for each item, despite the fact that some items had more than one mark available and therefore different chance probabilities for disagreement. Although the data is comparable across markers it may not be considered precisely comparable across items.

or where part of the student's response had to be drawn. The item level information contained in Tables 5.1.1A to 5.2.4A may be of particular interest to senior markers and the test development teams, who may wish to examine the characteristics of questions with poor between-marker levels of agreement.

In the main study, the mean percentage exact agreement (across markers) for new technology marking was 98 per cent for clerical markers and between 93 and 95 per cent for expert markers (Papers 1 and 2). However, as might be expected, there was a greater range of marking accuracy in the main study than the differences between the four markers in Study 5.

4.3.2.2 Analysis of within-marker reliability for conventional versus new technology marking

Within-marker reliability was examined by comparing the conventional and new technology marks awarded by each marker for the same scripts. Agreement between paper mark totals from the two systems was investigated using correlation statistics and mean mark differences. The percentage of items for which there was exact agreement between conventional and new technology marks was also computed, at item and paper level.

4.3.2.2.1 Correlation

The correlation between conventional and new technology marks for each marker, for each paper, is presented in Table 4.16.

Table 4.16 The correlation between conventional and new technology marks

Tier	Marker	Correlation	N	p
Paper 1, Tier 3 to 5	1	0.9971	25	0.000
	2	0.9957	25	0.000
	3	0.9956	25	0.000
	4	0.9955	25	0.000
	Overall	0.9955	100	0.000
Paper 2, Tier 4 to 6	1	0.9845	25	0.000
	2	0.9841	25	0.000
	3	0.9891	25	0.000
	4	0.9779	25	0.000
	Overall	0.9806	100	0.000
Paper 1, Tier 5 to 7	1	0.9750	25	0.000
	2	0.9816	25	0.000
	3	0.9815	25	0.000
	4	0.9805	25	0.000
	Overall	0.9749	100	0.000
Paper 2, Tier 6 to 8	1	0.9902	25	0.000
	2	0.9868	25	0.000
	3	0.9878	25	0.000
	4	0.9871	25	0.000
	Overall	0.9868	100	0.000

As with the between-marker correlations, the between system comparisons showed very high levels of association between conventional and new technology marks. As reported in the 2001 pilot report, the reliability correlation coefficients for mathematics papers are generally very high.

4.3.2.2.2 Exact agreement

Tables 5.3.1A to 5.3.4A show the extent of exact agreement between the two systems at both item and paper level, for each marker, and then across the four markers as a whole. Where there was disagreement between the new technology and conventional marks, these tables show the percentage of clips for which the new technology marks were greater than conventional ones and the percentage of clips where conventional marks were higher than e-marks.[23] A summary of the exact agreement data is shown in Table 4.17.

[23] As each marker marked 25 scripts in each tier, a figure of four per cent represents one clip.

Table 4.17 The mean percentage exact agreement between conventional and new technology marks within markers.

	Marker 1	Marker 2	Marker 3	Marker 4	All Markers
Paper 1, Tier 3 to 5	97.7	97.8	97.9	98.4	97.9
Paper 2, Tier 4 to 6	97.5	94.8	97.2	95.1	96.1
Paper 1, Tier 5 to 7	96.7	94.7	96.1	94.8	95.6
Paper 2, tier 6 to 8	96.5	93.7	96.7	95.5	95.6

Although there were some slight differences across markers, i.e., some markers were more reliable than others, the levels of exact agreement within marker were generally very high across all papers. The levels of agreement within marker were also very similar to the between marker reliability figures, shown in Table 4.14. As can be seen in Tables 5.3.1A to 5.3.4A, where there was within-marker disagreement, neither marking system led to a consistent bias in respect of the marks awarded, i.e. a general lenience or harshness. For each paper, the percentage of clips for which new technology marks were greater than conventional marks were very similar to the percentage for which the reverse was true.

4.3.2.2.3 Mean mark differences

As might be expected from the previous findings, the mean mark differences between conventional marking and new technology marks were extremely small (and statistically insignificant), as shown in Table 4.18.

Table 4.18 The mean mark difference between conventional and new technology marks.

	Mean mark difference (CM – NT)
Paper 1, Tier 3 to 5	0.003
Paper 2, Tier 4 to 6	-0.003
Paper 1, Tier 5 to 7	-0.002
Paper 2, tier 6 to 8	-0.0005

These findings are in sharp contrast to the 2001 pilot, when the new technology marks were found to be significantly lower than conventional marks across all the papers and subjects. Given that both the conventional and new technology marking were carried out by the same markers, the findings from this study suggest that the task demands of conventional and new technology marking are comparable and that moving to a new technology system would not impact on the student outcomes for mathematics. The between and within-marker reliability data would seem to provide evidence that mutually

validates the marking from conventional and new technology marking systems. However, this is a small study involving only 25 scripts from each tier and must be viewed in the context of data from the main 2002 pilot.

4.4 Summary

The aims of Study 4 were: to evaluate the wealth of data that it is possible to produce using e-marking technology; to compare the findings of the current pilot with those of 2001; to provide a body of data that can be used as a baseline for future pilots and trials; and to compare data arising from the pilot with conventional marking data. In addition, Study 5 provided a more direct comparison between conventional and new technology marking.

In collaboration with NCS Pearson, a large amount of management and measurement data from a variety of components of the new technology system were produced. In contrast to the lack of quantitative information from conventional systems, it is clear that e-marking enables the production of a vast amount of data on the speed, accuracy and reliability of its procedures. Such data has the potential to greatly enhance the effectiveness and defensibility of external marking and support managers, test developers, senior markers and researchers in making decisions as to how the system can be improved.

However, the sheer wealth of data available may also create new and significant problems for the external marking system. Firstly, there is a danger that there will be over reliance on the data produced and that inaccurate data, due to system error, may not be identified. There is also the possibility that managers and senior markers may be overwhelmed by the volume of data produced and that significant data may be lost or overlooked. Too much data or insufficient information as to how it can be used may therefore result in some of the potential uses of the new technology system being under-utilised. Finally, there are risks that the data will be misinterpreted or misused. The use of data for monitoring expert markers may pose specific risks and it is essential that managers and senior markers are given sufficient training as to how to interpret and use such information.

The potential benefits of new technology marking depend on the highest standards, both in terms of the marking technology and the provision of accurate and meaningful data. In the 2002, pilot deficiencies in the data production systems resulted in errors occurring in both the management and measurement data. Most of these errors were in the marker monitoring data rather than the student data (although one error was found in the data relating to students' conventional marks). Although these errors were identified and rectified during the data analyses, quality assurance must be significantly improved to avoid any possibility of such errors occurring in a scaled-up operation. Confidence in a

new system of marking would be severely jeopardised if there were any doubts concerning any of the data outcomes.

4.4.1 Issues arising from the management data

Information on the speed of the new technology system was collected in order to evaluate the extent to which marking could enhance the efficiency of the external marking process. Based on the findings of both the 2001 and 2002 pilots, the greatest threat would appear to be the potential bottleneck that could arise between the receipt of scripts from schools and the commencement of marking. In 2002, the level of exceptions and attachments resulted in a scanning rate less than half the targeted rate and a much slower turnaround of scripts than had been anticipated in Cycle 1. In a system employing new technology marking only, scanning would not have to be completed before marking commenced, but a sufficient flow to markers would need to be maintained. Although the level of exceptions would be expected to be reduced by redesigned test papers, further research into this issue may be required to establish the level of exceptions that would still remain, the means by which such rejected scripts could be efficiently processed and the time it would take to do so.

A further issue arising from the management data concerned the extent to which the new technology system can enhance the utilisation of human resources. Two of the major advantages of e-marking include using non-expert markers to mark data entry and clerical items and relieving expert markers of clerical and administrative tasks, such as the addition of marks, packaging scripts, etc.. The management data provides some important baseline information about the length of time taken to mark the different types of questions: data entry; clerical and expert. However, this needs to be contextualised by further information from NCS Pearson concerning the overall time taken to complete the marking by the various different types of marker. As reported in Section 2, the problems reported by home-based markers in respect of download speed will need to be addressed if a scaled up home-based operation is to be feasible. However, the productivity data does indicate the potential of new technology marking for making better use of the expert KS3 markers. It suggests that by eliminating the clerical tasks and reducing the number of items the experts are required to mark, the amount of time expert markers would spend per pupil would be approximately half that of conventional marking.

However, as the 2002 pilot has shown, there is a danger that with the double-marking of scripts the marking, supervision and adjudication demands will rise and any positive effects may be attenuated. Although less than one per cent of clips viewed by the expert markers were sent to supervisors unmarked, and only five per cent of expert items

required adjudication, it must be remembered that this was an extremely experienced group of expert markers, who had already completed their conventional marking and were very familiar with the 2002 mark scheme. It is likely that the number of items sent for review and those requiring adjudication would be much higher with a more representative marking population and an unfamiliar mark scheme (i.e. if new technology marking was the only marking taking place). The issue as to the purpose that double marking will serve and therefore which items will be double-marked needs to be fully debated. For further discussion of this issue see 5.1.4.

One finding of the 2002 pilot is the potential difference in the workload of supervisors working in a marking centre compared to those supervising home-based markers. Although the level of adjudication was similar, expert markers working from home sent more unmarked items to their supervisor than centre-based markers. Also, supporting the markers was more time consuming, as the home-based supervisor had to respond to large numbers of messages electronically. The implication is that in a scaled-up operation the supervision of home-based markers would need to be staffed more generously than a centre-based team.

4.4.2 Issues arising from the measurement data

Some important conclusions can be drawn from the measurement data. The correlation coefficients between new technology marks and conventional marks were very high (generally 0.98 or higher) and the mean mark differences were low and not statistically significant. Also the level of exact agreement between new technology and conventional markers, averaged across questions, was high for all the mathematics papers. These findings would suggest that, for mathematics, new technology marks are as valid as conventional marks and vice versa.

As the percentage of exact agreement between new technology marks and conventional marks was generally better for data entry questions and clerical items than for expert items, this finding supports that of the 2001 pilot, i.e. that using non-experts to mark data entry and clerical items can be technically effective. Further, that this can be achieved with a relatively low level of supervision. Generally, less than one per cent of data entry or clerical items were sent to supervisors without marking, and levels of adjudication ranged from one to three per cent.

On Papers 1 and 2 the between-marker reliability (levels of exact agreement between markers) was 98 per cent for clerical markers and between 93 and 95 per cent for expert markers. The lowest mean percentage exact agreement for an individual expert marker was 86 per cent. The between-marker reliability data therefore supports the findings from

the correlation and mean mark analyses and provides a wealth of baseline information for considering how marking quality could be evaluated in a future new technology system.

The absolute marker reliability data raises some important issues about the selection of validity items, the rate at which such items should be generated and how validity data should be utilised. The importance of this function largely depends upon the marking and supervisory models that are adopted in any scaled-up operation.

A significant advantage of a new technology marking system would be the elimination of the script annotation error that occurs within the conventional marking system and has been identified in both the 2001 and 2002 pilots. Based on the results of the 2002 pilot, the level of such error could be as high as four per cent of scripts for the main KS3 mathematics papers.

A key issue to resolve is how to make economical use of the monitoring data so that useful statistics can be calculated, which give an immediate indication of the fitness of markers. This study provides ample baseline data to inform such decisions.

4.4.3 Issues arising from Study 5

The findings from the Study 4 measurement data are further supported by those of Study 5: the controlled evaluation of marking reliability. In this small study, the correlation coefficients between new technology and conventional marks were also very high, as were the within system comparisons: the between-marker reliability; and the exact agreement analyses. The mean mark differences were insignificant, suggesting that the task demands of the two systems were comparable for mathematics marking.

All of these findings provide evidence that mutually validates the marking from conventional and new technology systems and suggests that the introduction of a new technology marking system for mathematics would not impact greatly or adversely on student outcomes.

5 An evaluation of the 2002 New Technologies Pilot

The preceding sections have presented the findings from the seven studies that together formed the evaluation of the New Technologies Pilot, 2002. This section presents a discussion of the evaluation objectives in the light of these results.

The four objectives of the evaluation were as follows:

1. to evaluate whether the NTP contractor managed successfully to implement the agreed procedures for 2002;
2. to evaluate whether the procedures implemented during 2002 were effective in delivering significant benefits without undue costs;[24]
3. to consider whether the procedures implemented during 2002 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.

5.1 Specific objectives for 2002

In addition to the general objectives outlined above, seven specific objectives, based on particular features of the 2002 pilot, were identified as principle foci for the evaluation:

1. to compare conventional and new technology marking standards;
2. to highlight significant marker, and marker training, issues arising from the 2002 e-marking models;
3. to compare strengths and weaknesses between centre-based and web-based e-marking models;
4. to explore the merits of double-marking;
5. to identify process scale-up issues;
6. to determine whether management information systems functioned optimally;
7. to evaluate the effectiveness of system improvements for 2002.

[24] Costs are to be understood as any negative consequence and not simply in financial terms.

Section 5 will attempt to summarise the findings from the seven studies, outline the lessons to be learnt and make some general conclusions and recommendations in response to the general and specific objectives of the 2002 evaluation.

5.1.1 Conventional versus new technology marking standards

In the 2002 pilot, the comparison of conventional and new technology marks in Studies 4 and 5 provided evidence that mutually validated both systems for mathematics tests. The findings also supported those of the 2001 pilot in the conclusion that the use of unskilled and semi-skilled markers could be technically effective for marking non-expert questions.

The differences in mean scores between conventional and new technology marking were not statistically significant. In Study 5, there were no differences in mean scores when the same markers marked the same scripts. In the main pilot, the mean mark differences were not statistically significant, yet were all in the same direction (i.e. the means for conventional marks were slightly higher than the equivalent new technology means). Although the data suggests that these results have occurred by chance, it is possible that BtCIA responses and the loss of the 'halo' effect are still resulting in marginally lower scores from new technology marking compared to conventional marking (see Section 4.2.3.2). Although the concordance between the levels achieved by students with conventional and new technology marks was extremely high, approximately four per cent of students would have been awarded a lower level and only one per cent a higher level if new technology marks had been used. Greater marker reliability may not necessarily equate to consistency with past standards. In the current educational climate, skilful change management will be required to reduce the risk that this issue may pose a threat to the scale-up of a new technology system. More detailed research focusing on this issue may be required.

Some issues of face validity, the extent to which new technology marking is perceived to be as fair as conventional marking, were raised by markers during the pilot. In addition to the problem of not being able to mark BtCIA responses, a further issue was the loss of contextual information to aid professional judgement (e.g. in deciphering handwriting) and to identify malpractice. (However, the loss of contextual information was also viewed positively in that it would reduce the 'halo' effect.) Although these issues were not viewed as major obstacles to the implementation of new technology, consideration should be given as to how they could be resolved, for example by changes to test design or improvements to the software. In the case of malpractice, it is possible that in the future the data generated by on-line marking may provide statistical means by which this could be identified.

5.1.2 Marker, and marker training, issues arising from the 2002 e-marking models

Although there is some concern regarding the extent to which markers in the pilot were representative of the marker population as a whole, the attitude of the participating markers towards on-line marking, as experienced during the pilot, was extremely positive. As reported in Section 2, some features of the software operated sub-optimally or were unsatisfactory (e.g. no recall function) but generally these were viewed by markers as technical issues to be resolved and did not cause dissatisfaction with on-line marking per se.

Although the attitude of the supervisors towards the system was also positive, they considered that their training had been inadequate and there had been insufficient attention to the planning and trialling of the procedures for supervision before the pilot went live. There were also failures in delivering some of the software requirements for training and monitoring markers, such as the training, pre-qualification and calibration modules. As a result of the unavailability of these modules, several potential aspects of the supervisor's role in a new technology system could not be evaluated. Although these aspects of functionality were not central to the double-marking model employed in the 2002 pilot, if the eventual marking model were to be essentially one of single-marking, these aspects would need to be fully piloted (see 5.1.4).

The demands on supervisors were found to be higher than had been anticipated. Given that markers in the pilot were relatively senior and had already completed their conventional marking, this suggests that the true workload for supervisors could be even higher. Although some of the 2002 workload is likely to be reduced once adjustments to the software have been made (e.g. if markers can self-correct errors with a recall facility), the workload implications in terms of human resources will need to be addressed.

Another issue arising from the pilot was that expert markers were aware that on-line marking opened up the possibility of a different payment model to that which operates within conventional marking. Supervisors were optimistic that the information derived from on-line marking could lead to a system of payment that rewarded accuracy and that this could have a positive effect on recruiting and retaining high quality markers. However, there were concerns that the allocation of work would need to be carefully controlled or linked to whatever payment model was implemented (for example markers should be given an equal number of 'easy' and 'difficult' items to mark or that payment should be linked to individual items). Such models would be far more complicated than

that currently used in conventional marking and consideration should be given to piloting how such payment models would operate.

5.1.3 Centre-based versus web-based e-marking processes

A major development within the 2002 pilot was the introduction of home/web-based expert marking. This was implemented alongside a parallel centre-based operation, where expert markers were brought together in a dedicated marking centre to mark across an intranet. The home-based markers were supported primarily on-line, although initial training was held in the marking centre. The centre-based markers received both training and on-going support face-to face. All non-expert marking was carried out in the marking centre.

The 2002 survey of external markers found that markers were generally against the idea of marking centres (even local ones) and that the introduction of such centres was thought likely to have a negative effect on recruitment and retention. Although positive about the possibility of marking from home via the internet, markers indicated a preference for face-to face training in order to benefit from shared discussions with colleagues and to develop a shared understanding of the mark scheme. In the main, these views would appear to be supported by the findings of the pilot. However, although the number of markers in the pilot was small, it is interesting to note that some of the expert markers who marked in the centre were far more positive about the experience than they had anticipated. They had enjoyed the opportunities afforded by the centre to discuss marking issues with colleagues and obtain speedy, face-to-face advice and support from their supervisor. Nevertheless, given that many markers work either part-time or full-time, it is unlikely that a fully-scaled up operation could succeed unless home-based marking was an efficient component of the system on offer.

Although the pilot demonstrated that home-based marking was both possible and, in some cases, successful, several home-based markers experienced difficulties (predominantly slow download speed) due to hardware or internet connection issues. However, information from the pilot will enable informed decisions about minimum specifications for efficient home-based on-line marking. Further research needs to be carried out as to the extent to which on-line marking would affect the recruitment and retention of KS3 mathematics markers and whether providing or subsidising markers to obtain or upgrade their home computer equipment would have a positive effect on recruitment.

The potential costs involved in centre based expert marking, i.e. accommodation and travelling expenses, could be higher than subsidising the costs of home-based marking. In the same way that many markers would be unable or unwilling to attend marking centres,

there could also be problems in recruiting sufficient supervisors. However, as the 2002 pilot demonstrated, the supervisory workload was somewhat higher for the home-based supervisor than the centre-based one and could therefore result in higher supervisory costs for a home-based operation. It is probable that, although all non-expert marking would be carried out in marking centres, a mixed centre-based and home-based operation would offer most flexibility for expert markers and supervisors.

For the 2002 pilot, all training (both mark scheme refresher training and e-PEN training) was carried out in the marking centre; there was no trial of home-based on-line training. One of the potential benefits of home-based marking is to reduce costs associated with training. Although markers in the pilot were strongly in favour of face-to face training in respect of the mark scheme, it is possible that software training and non-confidential administrative training could be completed on-line.

5.1.4 Merits of double-marking

In contrast to the 2001 pilot, in which a mixed marking model was employed, the 2002 pilot incorporated a 100 per cent double-marking model. The latter has much to recommend it in terms of improved marking reliability, due to the adjudication of unmatched mark pairs by senior markers (supervisors). In the pilot approximately four to five per cent of expert pairs of clips were adjudicated, suggesting that double-marking could potentially eliminate errors of approximately half that percentage (assuming that with single marking a marker might be correct in 50 per cent of cases). A further one or two per cent of expert clips were referred to supervisors to be marked, thus avoiding other potential marking errors. However, the potential savings, in terms of speed, costs and scarce resources (expert markers), achieved by reducing the number of items marked by experts may be lost if a double-marking model were to be adopted.

In the 2002 pilot, the marking generally took longer than had been anticipated and additional reserve markers had to be employed to complete the marking within the agreed dates. It therefore provides important baseline information about the numbers of markers and supervisors that would be required to complete 100 per cent double-marking of a set number of scripts within a specified period. The pilot also provides data in respect of marker reliability and the associated supervisory demands to inform decisions about the choice of marking model. New technology software offers considerable flexibility in respect of single or double-marking. For example, it would be possible to double-mark a specified proportion of items and then reduce to 75 or 50 per cent double-marking. Alternatively, decisions about the level of double-marking could be made on an item by item basis.

Some of the perceived deficiencies of the software and the monitoring data in 2002 were because they were designed to support quality control (within a double-marking model) rather than quality assurance (within a single-marking model). For example, there were insufficient validity items to enable valid inferences about quality of marking in sufficient time to influence marking on subsequent items (although they could be used for post-hoc accountability/auditing purposes). Also, messages sent by supervisors, in response to clips referred by markers, could not be returned with the original clip attached – therefore the provision of formative feedback was very limited. Double-marking increases accuracy but potentially increases supervisory load due to the emphasis on adjudication. The reduction in formative feedback increases the likelihood that markers will continue to make the same mistakes. This lack of feedback may also eventually lead to markers experiencing less ownership of their marking, possibly resulting in less careful marking.

The fundamental decision as to whether a double or single-marking model is adopted has implications for type of supervision and marking hierarchy required and which features of the marking software are most critical. It is therefore essential that in any future pilot consideration is given to the role of supervisors in relation to the marking model that is likely to be adopted in a fully scaled-up operation. Training for supervisors could then be tailored to the supervisory role that is required, i.e. is their role essentially one of quality control or quality assurance? Specifications in respect of software should also support the marking model adopted. If a predominantly single-marking model is adopted, the marking agency has to be confident that marking is to a sufficiently high standard. As in conventional marking, supervisors would be assessing competence at the outset and then supporting markers to maintain that standard. Therefore, pre-qualification (i.e. a 'certification' stage), calibration and validity functions would become much more important and rapid feedback would be crucial, particularly at the beginning of a marking period. Supervisors would need to be able to prioritise 'send to review' items, where markers require guidance on the mark scheme, and it would be essential to be able to return the clip itself with an appropriate message. Similarly, the rationale for technical decisions, such as the rate at which validity items are delivered to markers, should be based on the role that the validity function is required to perform. If supervisors are carrying out a quality assurance role (single-marking) rather than quality control (double-marking) then the validity rate needs to be much higher, so that they can monitor marking quality during the marking period and stop markers who are below the required standard.

5.1.5 Process scale-up issues

The major scale-up issue addressed within the 2002 pilot was the scanning of a much larger number of scripts than was attempted in 2001. As reported in Section 4, NCS Pearson successfully scanned approximately 36,000 of the 41,000 scripts received. Unfortunately the requirement not to have to scan problematic scripts in 2002 did not improve hourly scanning rates due to a high level of exceptions and the time taken to remove these from the scanners. As a result, the time taken to process scripts through the batching, slitting and scanning stages was much slower than the target turnaround of 48 hours. The findings from Study 7 suggest that most of the scanning exceptions occurred because students were obscuring the timing marks (added to enable scanning) and that, if test papers were redesigned in order to give students more defined areas in which to write responses, the level of exceptions would be likely to be considerably reduced. Redesigned test papers may also help to reduce the incidence of responses BtCIA. Changes to the design of test papers and their administration would need to be communicated and managed effectively. However, improvements to the speed of scanning and procedures for dealing with any remaining exceptions (including scripts from students with special arrangements) will have to be implemented if on-line marking of KS3 mathematics is to be scaled-up further.

As mentioned in 5.1.3, scaling-up may also be constrained by hardware issues in relation to home-based marking. Although the availability of markers with the necessary equipment that would enable them to mark on-line is likely to increase over time, in the short-term, recruitment of sufficient markers able to do so may be one of the potential barriers to a fully scaled-up operation. Apart from the specific issue of hardware, the 2002 survey of external markers suggested a substantial number of KS3 mathematics markers (30%) were currently unwilling to mark on-line. Dissemination of information about the nature of the on-line marking that is being trialled and the positive reactions of markers who have been involved in the 2002 pilot may well overcome some of this resistance. However, further research into the potential effects of the introduction of new technology marking on recruitment and retention would be advisable.

A further major issue is to decide whether a double-marking model is appropriate and feasible for the UK external marking system. It is important to take key decisions about the direction of any future pilot so that clear technical and human resource specifications can be made, as to what is required and by when, and that sufficient time is allowed for the planning, development and trialling of systems and procedures. Given the reliance on technology, both of the contractor and of the home-based markers, consideration should be given as to the risk assessments that should be undertaken, in order to specify the

disaster recovery and contingency plans that need to be in place prior to further expansion. As part of the risk assessment, health and safety guidelines should be produced in recognition of the likely impact on markers of the change from conventional to new technology marking.

5.1.6 Management information systems

One of the key features of new technology marking is that it allows the production of a wealth of data in addition to the student outcomes. Unfortunately, in the 2002 pilot the management data production did not appear to have received the same level of planning and quality assurance as the marking software or the processing and tracking of scripts through the system. A plethora of monitoring data was produced during the marking phase but many of these reports were not user-friendly and did not support the on-going monitoring of markers. Some of the management and measurement data produced after the marking phase was found to be incomplete or inaccurate. The quality of this data would be unacceptable in a scaled-up operation where such information might be used to support decisions about marker accuracy that in turn might be linked to marker remuneration.

Once a decision has been made as to whether a single-marking or double-marking model is to be adopted, more consideration should be given as to the type of information that needs to be provided, and at what stage. For example, what information should be provided during the live marking phase and what additional data is needed after the event to make decisions about marker performance? Monitoring data to be used during marking should be kept to a simple and manageable level so that supervisors are not overwhelmed. Also, they should be given specific training on what the data means and how to use it. The 2002 pilot provides data on productivity, reliability and validity to inform decisions about what indices of acceptable marking might be used to assess marking quality.

5.1.7 System improvements

Several changes to procedures and systems were implemented by the contractor for the 2002 pilot. Overall, the use of e-PEN rather than Netgrade software was viewed as a significant improvement by those markers and members of the evaluation team who had been involved in both the 2001 and 2002 pilots. However, some features available with Netgrade in 2001, such as the facility to recall marked items, were not included in the e-PEN software used for the 2002 pilot. Other features of the e-PEN system (e.g. the calibration module) that had been promised by the contractor for 2002 were not available in time for the pilot. Although the main features of the e-PEN system were observed and

evaluated, the evaluation was limited to the extent that the system employed in 2002 is not the fully developed version that would be used in a scaled-up operation.

One of the main system improvements was the development of the on-line mark scheme. Although some further minor changes were requested by markers, the main features of the on-line mark scheme were found to be satisfactory by most markers. Although some markers expressed a preference for their paper copy, in many cases this was because they had already made annotations during conventional marking and were familiar with its layout. Supervisors particularly liked the facility to make global annotations to the on-line mark scheme, for example being able to add examples that would clarify acceptable or unacceptable responses. However, given the difficulties of scrolling on screen, most markers felt that a hard copy of the mark scheme would always be necessary in addition to the on-line version.

In the 2001 pilot, there were insufficient expert markers with prior experience of conventional marking. In 2002, NCS Pearson, in conjunction with AQA, succeeded in recruiting sufficient experienced KS3 mathematics markers. These expert markers adjusted easily to the new technology marking system, found it easy to use and generally were very positive towards on-line marking. However, the sample of markers was again unrepresentative, in that there were proportionately more senior markers and team leaders in the pilot than in the general AQA marking population as a whole. It is difficult to estimate with any certainty how this has affected the evaluation.

5.2 Conclusions

5.2.1 To evaluate whether the NTP contractor managed successfully to implement the agreed procedures for 2002;

NCS Pearson was largely successful in repeating the key components of the 2001 pilot and implementing the agreed procedures of the 2002 pilot: scanning a much larger sample of scripts; using a full ability range key stage test; and extending the on-line marking to home-based markers. Whilst demonstrating the enormous potential of on-line marking for national curriculum tests, some procedures and/or aspects of system functionality were not implemented or were not wholly successful. Although these did not compromise the pilot, these areas of weakness would need to be addressed in the future.

A major issue that was not resolved by the 2002 pilot was the optimum scanning speed that could be achieved with a large volume of scripts. Although it would be unreasonable to expect the contractor to have anticipated the high level of exceptions that occurred, NCS Pearson must take responsibility for the failure to scan the majority of clips in bi-

tone rather than greyscale and the failure to anticipate the impact of the agreed clip sizes on scanning rates. The other areas of weakness were supervisor training, the failure to deliver some software components (e.g. pre-qualification/training, calibration), the variability with which web-based markers were able to mark effectively, and poor quality control with regard to the monitoring and management data. There were also some procedural concerns, i.e. that some operational decisions (for example the validity rate and changes to the allocation of questions to marker types) were taken for reasons of expediency and with insufficient consultation.

5.2.2 To evaluate whether the procedures implemented during 2002 were effective in delivering significant benefits without undue costs;

In a new technology system such as that piloted in 2002, expert markers would be relieved of all the clerical and administrative tasks currently associated with conventional marking. Although the time gained in the elimination of certain conventional stages (adding marks, completing mark sheets, data input, etc) is likely to be lost by the introduction of new technology ones (batching, slitting, and scanning), these stages would not be utilising scarce human resources (KS3 mathematics markers). Using unskilled and semi-skilled markers to mark non-expert items has been shown to be effective and would also result in a more efficient use of marker expertise. The evidence from the pilot does not suggest that on-line marking would be more rapid than conventional marking. Also, a new technology system would not necessarily result in financial savings (it is likely that in the short-term at least costs could well increase). However, the on-line marking model employed in the 2002 pilot has shown that there could be considerable benefits, including enhanced marking reliability, elimination of clerical errors and the provision of detailed information for managers, test developers, and participating schools.

The 2002 pilot has demonstrated that a mixed centre-based and web-based on-line marking model has the potential to increase the accuracy of the external marking system, providing the following can be achieved:

- sufficient unskilled and semi-skilled markers could be recruited during the marking period to mark non-expert items;
- sufficient expert markers have (or could be provided with) the necessary equipment to optimise home-based marking or could be persuaded to mark in marking centres;

- the modifications requested by markers and supervisors can be made to the e-PEN system (or equivalent).

Unfortunately, it is not possible to predict the likely benefits or risks in terms of the speed of on-line marking until the processing of scripts is piloted with test papers optimised for use with high speed scanners and the procedures and time required for dealing with all exceptions/attachments are fully investigated.

5.2.3 To consider whether the procedures implemented during 2002 might be scaled-up for all national curriculum tests to deliver significant benefits without undue costs;

Given that the expert items in each script took approximately twice as long to mark as the non-expert items, a double-marking model is likely to require as many expert markers as a conventional system and would be likely to create a heavier supervisory workload. Consideration must therefore be given to the number of expert markers and supervisors that would be required to implement a double-marking model and whether this is feasible with the current rates of recruitment and retention of KS3 mathematics markers. If such a model is adopted, the role of supervisors needs to be more clearly defined than was evident in the 2002 pilot. It appeared that the supervisors experienced some problems in adapting their conventional supervisory practice to the double-marking model and found it difficult to achieve a satisfactory balance between quality control and quality assurance. If double-marking is to be adopted, with an emphasis on quality control, it would be beneficial to investigate several different models of supervision to identify the most effective one. On the other hand, if it were decided to employ a single-marking model, it would be necessary for the software to support a supervisory model focusing more on quality assurance (as outlined in 5.1.4). Marking hierarchies may also need to change to accommodate new supervisory practices.

Generally speaking, the centre-based marking component of the 2002 pilot was extremely successful. Scaling-up to a national level would depend on sufficient non-experts being recruited at the appropriate time and sufficient expert markers being available and willing to attend dedicated marking centres. Given that many expert markers are full-time or part-time teachers, the number who would be able to do this may be limited. A centre-based model would also introduce additional costs relating to the set-up and maintenance of marking centres and the travel and accommodation expenses of markers. Scaling-up the web-based operation would offer considerable benefits in terms of costs. However, this may not be feasible in the short- to medium-term, unless markers are offered considerable inducements to up-grade their computer hardware and internet connections.

As in the 2002 pilot, a major risk is that, in using markers who have less than the optimal hardware required, they will be dissatisfied with the speed of on-line marking and therefore the associated remuneration, and that this will impact negatively on marker retention. However, an interim solution may be to investigate the possibility of markers using school-based computer facilities during evenings, weekends and school holidays. If small groups of markers could use the same facilities locally, such groups could take advantage of some of the perceived benefits of a marking centre (discussion with colleagues and face to face supervision), while keeping operational costs low.

Although a large number of markers were interested in participating in the 2002 pilot, as has already been mentioned, the external marker survey indicates that other markers are less enthusiastic about on-line marking. Dissemination of information about on-line marking and further monitoring of marker attitudes and capabilities may be essential to facilitate a scale-up to a national level.

Teachers and schools would be likely to benefit considerably from a national scale-up of on-line marking, in that a new technology system would decrease the administrative burden and have the potential to provide schools with valuable information, such as student, class, question and topic area analyses of their results. There is of course a risk that the use of unskilled and semi-skilled markers in an on-line system could be perceived as a down-grading of the marking process. However, acceptance of on-line marking amongst teachers, parents and the general public is likely to be achieved if they can be convinced that improvements in the accuracy of marking are not at the expense of validity. Skilful management of the change from conventional to new technology marking will be required to ensure that confidence in the external marking system is maintained or strengthened.

New technology offers potential advantages in providing information to review and monitor all the various stages within the external marking process. In theory, an on-line system could therefore contribute to an improvement in the management of the external marking system. However, as shown in the 2002 pilot, there are considerable risks associated with large data production systems. The data provided by a new technology system may potentially be used to support functions that have not previously been possible (e.g. monitoring marker reliability). It is therefore essential that such information is accurately generated and that detailed planning and technical advice (from senior markers and experts in psychometrics) should determine the type of data that is produced and the uses to which it is put.

The findings from the 2002 pilot relate specifically to KS3 mathematics. As these generally support findings relating to the Y7 mathematics tests in the 2001 pilot, it would be reasonable to conclude that the procedures implemented in 2002 would produce similar results for other mathematics tests. However, as there were considerable differences in the findings for English and mathematics in the 2001 pilot, the extent to which the 2002 results can be generalised to other subject areas remains in doubt.

5.2.4 To consider whether revised procedures for future years might deliver significant benefits without undue costs.

The benefits of new technology systems in terms of enhanced efficiency and accuracy must be balanced by an awareness of the enormous risks associated with such a fundamental change in the external marking system. One of the over-arching concerns of the evaluation team has been the tension, demonstrated during the 2002 pilot, between the underlying philosophy of a technology-driven, private sector company (developing systems in an iterative fashion) and that of the educational organisations responsible for external marking within the UK (with public expectations of perfect, error-free systems). Any future implementation of a national on-line marking system will take place within a public arena, in which every facet will be scrutinised by the relevant stake-holders, the media and the wider general public. This cautions against a hurried implementation and emphasises the need for thorough testing, not only of the software itself, but also the underlying systems. In other words, there should be a full and careful specification of the ergonomics of the process and a full system test prior to use. In the 2002 pilot, there were several instances of errors in the data, due to procedural inadequacies, that would be unacceptable in a live operation.

One of the supposed benefits of a new technology system is the potential for opening up sectors within the external marking system to a number of suppliers and thus driving down costs by the introduction of competition. However, one of the lessons of the 2002 pilot was that whereas the marking software operated relatively successfully, several errors occurred when data was being transferred from one data base to another (e.g. transferring the PECS data and the e-PEN data into files suitable for transmission to the evaluation team). Such risks would be likely to increase where several contractors were handling different components of a system and information was being passed from one to another. Also, in the short-term at least, it is doubtful whether there would be a sufficient number of competent contractors to allow market forces to operate effectively.

5.3 Recommendations

Although considerable progress has been made during 2002, the system is not yet ready to be introduced formally for KS3 mathematics at national level.

1 Consideration should be given as to whether the marking model to be adopted will be a double-marking or single-marking plus sampling model, if a fully scaled-up operation is to be contemplated.

2 Depending on the marking model to be employed, specifications will need to be developed for modifications and additions to the software, in terms of the marking, supervisory and monitoring functions required. Also, training for both markers and supervisors will need to be tailored to support the particular marking model being used.

3 Formal plans should be developed as to what management and measurement data will be produced, how the quality of such data will be controlled and how such data will be used, (for example what indices of marker reliability will be applied). The data should be specific and limited to the particular use for which it is required.

4 Specifications should also be developed that reflect the on-line marking operation as a whole, detailing the procedures and systems involved at each stage, how they will be managed and who will be responsible for ensuring their accuracy.

5 Dissemination of information about current developments and future plans with regard to on-line marking should be carried out with external markers. Following such publicity, a detailed survey should be conducted into current marking practice and the effect on marker availability of implementing a new technology system, (including information about markers' computer systems and internet capabilities).

6 A minimum technical specification of the equipment required for effective and efficient home-based marking should be developed.

7 The possibility of expert markers using local marking centres, utilising school-based computer suites, for on-line marking should be investigated.

8 Research should be undertaken into models of work allocation and marker payment.

9 Further piloting for the new technology programme should be undertaken to test specific aspects that need to be examined in more depth. For example large-scale scanning using specifically designed test papers. Also, further investigation of the scanning resolution issue to identify the optimal settings required to maintain image quality, maximise scanner throughput and yet minimise the size of image files for transmission to home-based markers.

10 Work should be undertaken to establish what effective scanning speed is required in order to fit on-line marking into the external marking time frame.

11 Risk assessments should be undertaken to inform the development of disaster recovery plans and health and safety guidelines.

12 Consideration should be given as to whether further research be carried out, looking in detail at the differences between new technology and conventional marks. (For example, if copies of a small sample of the conventionally marked scripts from the 2002 pilot could be obtained back from participating schools, it would be possible to examine those items where mark differences occurred in order to identify the reasons why the new technology marks were lower than those produced conventionally.)

13 Separate research should be considered as to the viability of introducing on-line marking for other subject areas.

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 5DG.
Tel: 01904 433435
Fax: 01904 433436
E-mail: j.harland@nfer.ac.uk