Qualifications and Curriculum Authority

Interim evaluation of the 2005 pilot of the Key Stage 3 ICT tests

A report to the Department for Education and Skills

July 2005

QCA/06/2340

Preface

The Qualifications and Curriculum Authority (QCA) is working under contract to the Department for Education and Skills (DfES) to develop an on-screen test of Information and Communication Technology (ICT) at Key Stage 3. Subject to successful pilot, this test will become a statutory National Curriculum test by 2008.

Yearly pilots are informing the development of the ICT test, and this report evaluates the 2005 pilot. QCA commissioned Andrew Boyle, a researcher in e-assessment, to carry out the evaluation. He is employed by QCA as a researcher rather than a member of the team developing the test and the report is, therefore, independent.

This report was delivered to the DfES by the QCA on 12th July 2005. It evaluated the 2005 pilot's success against a set of objectives. Its findings reflected the fact that not all information about the 2005 pilot was available by that date. As such, this was an interim evaluation. It is supplemented by a final evaluation report.

Andrew Boyle
QCA
02 February 2006

Contents

List of tables

List of figures

1 Executive summary

This is an interim evaluation of the 2005 pilot of the Key Stage 3 ICT tests. This interim report will be followed by a final evaluation in September 2005.

The 2005 phase of the project culminated in a summative test pilot in April and May 2005. In that pilot, 45,540 pupils in 402 schools sent valid test data to a central server. National Curriculum levels in ICT have been set, and it is intended that the levels will be returned to these pupils and their teachers in the next few weeks.

This interim report evaluates the pilot's success with respect to five objectives. It divides evaluation outcomes into three categories:

- Has already been achieved
- Insufficient evidence yet
- Likely not to be achieved

For objectives (or aspects of objectives) that fall into the last two categories, a final evaluation will be made in the September report.

The following table summarises the interim findings:

Objective number	Objective focus	Sub-focus	Interim finding
1	Validity	N/A	Insufficient evidence yet
2	Infrastructure software and support processes scalability	Infrastructure software reliability and scalability	Has already been achieved
		Support processes scalability	Insufficient evidence yet
3	Accurate formative and summative reports	Summative reports	Insufficient evidence yet
		Formative reports	Likely not to be achieved
4	Test security	Classroom issues	Insufficient evidence yet
		Institutional issues	Has already been achieved
5	School experience	N/A	Has already been achieved

Table 1: Summary of interim findings for each objective

Additional findings and recommendations are summarised in the following list. The findings and recommendations are organised under 'positive outcomes' and 'areas for further work' sub-headings.

Objective one: Validity

Positive outcomes

- RM implemented, and QCA scrutinised, a high-quality programme of formal validation work.
- Thresholds for level awards were set at meetings that took place in the week beginning 27th June 2005.

Areas for further work

- Test development contained many quality assurance steps, but there were some important ways in which procedures could be improved.
- There were approximately one-and-a-half times as many opportunities addressing level five in the upper tier of the test as in the lower. The impact of the different number of opportunities addressing the shared level five in the two tiers should be investigated.
- There was an insufficient amount of level six material in the test.
- Level six material related almost entirely to the curriculum area of data handling. This had the potential to disadvantage pupils who had level six ICT capability overall, but were not as good at data handling.
- The basis for awarding level six was different to that for the lower levels. This was a function of the small amount of level six material in the test.
- Pupils were awarded level six if they were 'a firm level five', and if they had demonstrated at least one piece of level six evidence.
- Task instructions were felt by pupils, teachers and informed observers to be difficult to comprehend. This was a complex issue and may have had several causes (including issues in areas that were the responsibility of other programme strands). However, this issue could decrease validity if comprehension was a significant source of difficulty in the test, or if the impact of these problems was disproportionate on low-level pupils.
- Some pupils had problems working with several applications (cluttered screens, difficulties with the 'split screen' functionality). Software development should make the operating system more useable, but the issue might also depend on the extent to which pupils have been taught to work with multiple applications.

Objective two: Infrastructure software and support processes scalability

Positive outcomes

- The error rate in software coding was low, surpassing industry standards.

- The infrastructure software was straightforward for schools to install.
- Although the summative pilot materials were known to contain eight defects with priorities of 'urgent' or 'very high' at release, these were felt to be unlikely to occur in schools. The software was much more robust than that released in 2004.
- Almost 46,000 data sets were returned during the pilot. The large data return demonstrates the technical scalability of the software solution.
- There were no reported problems in maintaining CPS server connectivity to schools throughout the 2005 summative pilot window.

Areas for further work

- Software defects are becoming increasingly obscure and difficult to fix. The project may need to weigh up the benefits of fixing all defects, however obscure, against the potential expenditure of large amounts of resource to fix infrequently-occurring faults.

Objective three: Accurate formative and summative reports

Positive outcomes

- The QCA has recognised that the formative aspect of the project has not been prioritised to date, and has escalated formative reporting to an issues log.

Areas for further work

- Formative reports had a very low approval rating from teachers, in an initial collation and analysis of responses to a questionnaire item.
- Most teachers who gave a written response about formative reports said that they had not seen them; those who had seen them were almost all critical of them.
- The reasons for problems with formative reporting may come from a range of sources, one of which is beyond the remit of the test development project.

Objective four: Test security

Positive outcomes

- Initial impressions of the full summative pilot were that the issue of exam conditions was being taken seriously.
- In 2005 the independent consultant KPMG replicated penetration tests and a web-hosting infrastructure review (having done these same tests in 2004). They found that lessons had been learned and that there were far fewer vulnerabilities exposing the KS3 ICT system to risk in 2005 than there had been in 2004.

- The QCA project team is building up information security incrementally until the 2008 high-stakes administration of the tests. It believes that this approach embodies an active and informed approach to managing risk.

Areas for further work

- The majority of a small number of observed summative pre-test sessions were not being run under exam conditions.
- New KPMG reviews conducted in 2005 on security policies and procedures, and test development and management suggested that substantial work was needed to ensure the security and confidentiality of commercially or personally confidential information.
- KPMG commented that there was substantial work to be done to ensure security of information across the RM consortium. In particular, 3T seemed to have a substantial number of vulnerabilities.

Objective five: School experience

Positive outcomes

- Analysis of questionnaire responses in respect of familiarisation, practice and pre-test materials showed teachers' and pupils' positive experiences in respect of several aspects of the pilot.

Areas for further work

- Only about 400 schools sent valid data back from the summative pilot. This was fewer than the original target of between 500 and 600 schools. The reasons why the number of schools was low were not clear.
- It has been suggested that a Regulatory Impact Assessment (RIA) should be carried out before the Key Stage 3 ICT tests become a statutory requirement.
- Less favourable questionnaire findings for the project included:
  - The database application's relatively low approval ratings.
  - A majority of pupils not finding the pre-test interesting, and a majority not having enough time to finish it.
- Qualitative findings with respect to objective five included:
  - Concerns about fitting two pre-test sessions into school timetables.
  - Some evidence that more than one member of staff was being used to invigilate a session.
  - That it could be difficult for schools when one or two pupils only could not complete a session.
  - Concerns that lack of IT skills amongst some Exams Officers prevented them from carrying out administrative duties in relation to these tests.

2 Introduction

2.1 Purpose and scope of this report

1. This is a report evaluating the 2005 pilot of the Key Stage 3 Information and Communication Technology (ICT) tests.
2. This is an interim report in that it considers all information that is available by 12th July 2005. This deadline means that some important aspects of information are not available to the report.
3. This report makes full use of the broad range of sources of evidence that are available at the time of writing. A formal list of these sources is given in the evidence section on page 6. This broad range of sources means that different types of information are used: the output of large, formal data collections, and of smaller studies. Whilst both large and small studies inform this evaluation, only information that is believed to be reliable is used.
4. The fact that not all information on the pilot is available in time for the writing of this report means that this interim document will only make a categorical indication where an objective has already been achieved.
5. The current report will be followed in September 2005 by a report which will make final judgements about the project's success in achieving all its objectives.

2.2 Structure of this report

6. The report starts with a section that outlines the types of information upon which it is based. Then there is a section evaluating each individual objective. At the end of the report there are two annexes:
   - Annex A (page 34) gives some broad background to the Key Stage 3 ICT programme, test development project, and the current pilot.
   - Annex B (page 40) acknowledges colleagues who have contributed to the report by providing information or vetting comments.

3 Evidence

7. This report is based on a variety of sources and types of information. These include:

   - products provided by RM to QCA in pursuance of the former's contractual obligations (for example: specifications, release recommendations, reports, and so on)
   - reports of monitoring and evaluation work carried out by QCA staff (e.g. write-ups of visits to schools, minutes of stakeholder groups, other research work)
   - the QCA project team's release recommendations for versions of the test software or other products
   - trust management reports commissioned by QCA to review the approach to the handling of secure information throughout the project
   - specially convened meetings with members of QCA staff to probe evaluation issues
   - any other relevant and reliable information

4 Evaluation of specific objectives

4.1 Objective one

8. Objective one is:

   Develop and administer Key Stage 3 (KS3) ICT tests that will deliver a valid and reliable assessment of pupil performance and award defensible national curriculum levels 3 – 6.

9. Its associated Critical Success Factor is:

   Validity

   This CSF is met if the validity report produced by the RM consortium provides sufficient evidence to demonstrate to the QCA and DfES that the test is a valid and reliable assessment allowing the award of Levels 3 – 6 to those completing the test.

10. Other Success Factors associated with objective one are:

- All pupils completing the 2005 pilot test receive a National Curriculum level and supporting summative report.
- The majority of pilot schools and other stakeholders consider the test a valid assessment of the ICT Programme of Study.
- Robust statistical analyses support defensible National Curriculum levels 3 – 6.
- DfES receives summative levels for all pupils taking the test.

4.1.1 Findings

11. Several sources of data that will be necessary for Success Factors listed above are not available at the time of writing this interim report. Specifically:

- Level setting has recently been conducted, and pupils have not yet received their levels.
- The opinions of all participants in the 2005 pilot have not yet been fully collated and interpreted.
- The QCA has not yet sent the DfES national data sets showing summative levels.

12. It thus follows that the final evaluation report will contain important findings on validity that cannot be made in this interim document.
13. Validity is probably the central concept in ascertaining a test's fitness for purpose. It has had many definitions, and there have been many debates as to the concept's true meaning. This document is based on a pragmatic conception of validity, which is set out in the following statements:

- A test is more likely to be valid if its development can be shown to have followed best practice.

- Much can be learned about a test's validity from studying the test forms in themselves – i.e. before studying stakeholders' opinions about the tests, or analyses of data derived from their administration.
- Stakeholders' opinions about the functioning of the test are a crucial indicator of the test's validity.
- Formal statistical analyses must be produced to show that a test is a robust and appropriate measurement instrument.

14. The rest of this 'findings' sub-section is organised to reflect the principles contained in the statements above.

4.1.1.1 Test development procedures

15. Test development was carried out by RM. Quality assurance steps for 2005 development were set out in a specification document, the content of which was agreed by QCA. Further, QCA believe that actual development work was, by and large, carried out in accordance with the specification.
16. The specified and actually enacted test development procedures contained many steps that were appropriate for a high-quality test development exercise. However, there were a number of ways in which this process could be improved.
17. Below is a list of improvements that could be made to test development processes:

    - Increase the number of tasks that are commissioned, so that there is never a possibility of poor quality tasks getting into a test form because there is no alternative.
    - Review the training (and consequent skills) of task writers in the craft of writing test questions (as opposed to curriculum knowledge).
    - Make sure that, when RM sends tasks to QCA for acceptance testing, they have already been thoroughly quality assured. As such, acceptance testing work should not be undermined by the need to do quality assurance.
    - Make sure that all participants in test development (RM and QCA) have a joint understanding of timetables.
    - Improve the system for due diligence checking agreed amendments to tasks.
    - Conduct pre-test review meetings, in which suitable experts review the materials that have been pre-tested, alongside statistics from the pre-test.
    - Have an expert review each complete test form as an entity, in addition to reviewing tasks as discrete entities.
    - Develop a system for literal proof-reading the whole test system (task instructions, assets, data files, etc.) within the test software, and as the last step before releasing the test to schools.

18. The recommended improvements are different in type:

    - Some are straightforward to implement (e.g. making timetables clear to all).
    - Some are straightforward conceptually, but will be more difficult to implement, because of the generally compacted timescales of the current project (e.g. commissioning more tasks, conducting pre-test review meetings).

- Some will require novel approaches to create analogues to systems from pencil-and-paper testing (e.g. proof-reading all files within the virtual world on-screen).

19. It thus follows that not all the suggested improvements will be able to be implemented immediately, and also that significant new thinking will be necessary. Notwithstanding this, it remains important that test development procedures are of the highest quality.

4.1.1.2 The developed test forms

20. As described in paragraph 201 below, the 2005 tests were realised as two tiers (levels 3 – 5 and 4 – 6). It is useful (although, as the following paragraphs show, not unproblematic) to describe the amount of material in each test tier in terms of 'opportunities'. (See paragraphs 203 to 208 for definitions of opportunities and related Rules Base concepts.)
21. The numbers of opportunities in the two test forms were described in the key RM specification document prior to release of the 2005 tests as follows:

Level 3 to 5 Test	
Level	Total
3	93
4	85
5	96
Grand Total	274

Level 4 to 6 Test	
Level	Total
4	92
5	150
6	36
Grand Total	278

Table 2: Numbers of opportunities in the lower and upper tiers

22. Before discussing findings from these tables, it is useful to set out some general principles and good practice:

- In high-quality tests (such as National Curriculum tests), it is conventional to describe the 'numbers of available marks'. In NC test regulations, such descriptions are highly specific and prescriptive.
- In any standard test it would be important that different test tiers that awarded the same level made awards based on the same weight of evidence. The most natural way to achieve the same weight of evidence would be to have the same number of questions or marks addressed at the shared level. If this ideal situation did not pertain, then a scoring adjustment to compensate for any discrepancy in the number of questions would need to be put in place.

- Opportunities are a novel concept in measurement; their properties are not well established. As such, it may be that it does not matter how many opportunities are in a test overall. But equally, the number of opportunities in a test overall (or at a shared level in different tiers) might have a crucial impact on the test's measurement properties.

23. Following from these general principles, some findings with respect to level five material in the two tiers can be stated:

    - There were many more opportunities targeted at level five in the upper tier (150) than there were targeted at the same level in the lower tier (96).
    - This difference in the number of opportunities at level five might have been an artefact of some features of opportunities and elaborations (e.g. the presence of level six evidence in the upper tier might increase the number of level five opportunities, the presence of some opportunities related to an open-text question might also skew things).

24. Two further emerging findings can be stated:

    - At the recent awarding meeting, RM's measurement expert stated that he believed that his initial analyses showed the tests to be highly comparable.
    - In RM's proposal for 2006 tests, the numbers of opportunities addressing level five in the different tiers are more closely comparable.

25. Thus, it is not clear what the impact of the different numbers of level five opportunities in the two tiers is. However, it is clear that opportunities are a crucial component in the Key Stage 3 ICT tests model, and that their properties should be understood more clearly. An important initial action would be to investigate the impact of the different number of opportunities addressing the shared level five in the two tiers.
26. There are also important findings from the table that relate to level six. There were only 36 opportunities in the upper tier that addressed level six. It has been stated that this was because the level six opportunities covered approximately 75 per cent of the Rules Base elaborations for level six.
27. Whether the reason for the small amount of level six material in the test was due to the paucity of opportunities, or of elaborations, seems a rather dry debate. It is more important to examine the consequences of the small amount of level six material. The most important consequence of the small amount of material at level six was that the level setting meeting awarded level six on a different basis from other levels; level six was awarded on the basis that a pupil was a 'firm level five' and had demonstrated at least one piece of level six evidence (see paragraph 53 below for details of this decision).
28. The fact that the small amount of level six material in the test might be a significant issue was backed up by a finding of the RM pre-test report, which noted that there was only a small amount of level six activity in test data files.

Also, a group of expert secondary strategy consultants convened by RM expressed concern, in a meeting and questionnaire responses, as to whether the test they had been shown could properly permit pupils to demonstrate level six ICT capability.

29. A further concern about level six content in the 2005 test was that the vast majority of level six material was in the last task in the test, which related to data handling only. This had the potential to reduce the fairness of the tests. For instance, some pupils who might legitimately be assessed as level six overall might not be so good at data handling. In the 2005 test, it is arguable that such pupils would not have a fair chance to demonstrate their level six capabilities.
30. RM specification documents described the tests' coverage of the National Curriculum in terms of aspects of the ICT Programme of Study. The curriculum coverage was acceptable across the different aspects (with the exception of the concerns about level six content focusing exclusively on data handling).

4.1.1.3 Qualitative findings from school visits

31. Throughout the different phases of the 2005 pilot, schools were visited to find out whether the Key Stage 3 ICT testing system was suitable for its users. In the following paragraphs, several widely-observed issues relating to the tests' validity are described.
32. Substantial numbers of pupils, teachers and informed observers felt that task instructions could be difficult to understand. These groups of people put forward several possible reasons for instructions being hard to comprehend:
    - There was simply too much text in emails that conveyed instructions.
    - The vocabulary used in instructions was too difficult for some pupils.
    - The design of the email applet accentuated some of these problems. In particular, when the email applet was tiled with another open application, the message pane (containing the text of the instruction) could become extremely narrow.
33. There was particular concern that problems of comprehending task instructions were especially acute for less able pupils.
34. The concerns about task instructions from test stakeholders occurred despite the fact that RM had, in test development, sought to analyse instructions' readability using indices that are sometimes used to 'demonstrate' the readability of on-paper texts.
35. The problem of task instructions' comprehensibility is complex, and difficult to unpick. It probably depends upon a complex interaction of factors such as:

conventional (mode-independent) readability, screen design, and pupils' cognitive load whilst multi-tasking to solve a problem.

36. Pupils' difficulty in understanding task instructions may also have causes in areas that are the responsibility of the programme strand that engages with ICT teaching. For example, it may be that children tend not to read rubrics whatever the test, or it may also be that ICT teachers tend to deliver task instructions in class orally – thus making pupils less experienced at reading written instructions.
37. Despite these mitigating factors, some aspects of the work on comprehensibility of instructions have not, to date, been adequate. The Test Development Agency's reliance on conventional readability indices as the sole indicator of instructions' comprehensibility is not tenable; these indices are known to be of limited use with respect to any test rubric, and they also make no allowance for the fact that pupils are reading from screen, rather than from paper.
38. In general, it must be emphasised that this is a test of ICT, not of comprehension. If comprehension is a significant source of difficultly for certain pupils (or groups of pupils), then the test's validity will be seriously compromised.
39. Several expert observers and teachers remarked that the screen within the virtual toolbox could become very cluttered. This was particularly the case when a pupil had several applications open at once.
40. Although the virtual operating system would only permit two windows to be viewed at once, these windows could contain several panes, some of which could not be minimised or closed. The following figure is a rough sketch of the cluttered screen that a pupil in a pre-test session was observed to be using:

<table>
<tr><td rowspan="4">Email structure tree</td><td>Inbox</td><td rowspan="4">Clip art list (held as thumbnail graphics partially behind presentation thumbnails)</td><td rowspan="2">Presentation thumbnail</td><td rowspan="4">Full screen presentation slide</td></tr>
<tr><td>Email header</td></tr>
<tr><td>Attachment</td><td>Presentation thumbnail</td></tr>
<tr><td>Email message text</td><td>Presentation thumbnail</td></tr>
</table>

Figure 1: A sketch of a cluttered screen

41. Many pupils and teachers disliked the 'split screen' functionality that the operating system used to allocate windows to either side of the screen. Whilst some pupils

found this application useful (or at least said that they had got used to it), many found it difficult to use.

42. Improvements to the functionality of the virtual operating system and toolkit will hopefully make pupils' experience of the test more comfortable in 2006 and beyond.
43. However, it is possible that the underlying problem here is one of pupils who are not accustomed to a way of working in which problems are set, and the solution requires the use of several applications. This may imply that certain teaching techniques that are advocated in the secondary strategy still need to bed down.

4.1.1.4 Formal validation work

44. RM's formal validation work was thorough. This area of work was communicated to QCA in several precise and well-constructed specification documents.
45. QCA scrutinised RM validation work via written comments on specifications and in meetings of the Assessment Working Group.
46. RM's work addressed a wide range of facets of the construct of validity and was based upon two major data collections (a summative pre-test and the full summative pilot) and several subsidiary data gathering exercises (such as those to check whether pupils could copy by looking at their neighbours' screens, or test-retest reliability studies).
47. In accordance with a contract condition, RM designed a test system that awarded NC levels with the aid of the QCA Rules Base (defined in paragraph 203 of this report). In order to do this, RM developed and used a new approach to measurement.
48. The new measurement model was known as the 'sufficient evidence model'. The definition of the sufficient evidence model can be summarised as follows: pupils' actions in the test were captured, and then were aggregated to construct meaningful 'chunks' of ICT evidence.
49. Once the evidence that pupils had demonstrated during the test had been assembled into meaningful chunks, it was then 'weighed'. Finally, a judgement was made as to whether the amount of evidence with respect to each NC level was sufficient for the pupil to be awarded that level.
50. Level setting took place at two meetings in the week starting 27th June 2005. Levels were set according to the number of opportunities (defined in paragraph 207) that pupils had achieved at levels. The exact numbers are shown in the following table:

Tier	Level awarded	Threshold numbers of opportunities required for pupils to achieve levels
3-5	'N'	< 12 level 3, 4, 5 opportunities
	3	≥ 12 level 3, 4, 5 opportunities
	4	(≥ 11 level 4, 5 opportunities) AND (≥ 12 level 3, 4, 5 opportunities)
	5	(≥ 8 level 5 opportunities) AND (≥ 11 level 4, 5 opportunities)
4-6	'N'	< 15 level 4, 5, 6 opportunities
	4	≥ 15 level 4, 5, 6 opportunities
	5	(≥ 8 level 5, 6 opportunities) AND (≥ 15 level 4, 5, 6 opportunities)
	6	(≥1 level 6 opportunity) AND (≥16 level 5, 6 opportunities)

Table 3: Threshold numbers of opportunities agreed at awarding

51. Significant issues arising from this table include:

- Levels were awarded on the basis of a number of opportunities at a level.
- Awards for levels that were not the lowest in a tier required pupils to have achieved the same number of opportunities at the tier below the awarded level, as well as a specified number of opportunities at the level that they would be awarded. (For instance, in order to be awarded level four, a pupil would also have needed to have fired enough opportunities to be awarded level three.)
- The way in which level six was awarded was different to the way in which the other levels were awarded. The level six award was made on the basis that pupils were 'a firm level five' (i.e. had achieved well above the level five threshold) and had displayed a small amount of level six evidence (at least one opportunity).

52. The fact that level six had to be awarded on a different basis from the other levels was a function of the small amount of top-level material in the test (see paragraph 26 above).
53. The dilemma facing the awarding meeting was that to stick to the same awarding criteria as for the other levels would have meant that a very small proportion of pupils would have been awarded level six. This might have been considered unfair to pupils. However, to award level six on the basis of only one level six opportunity could have led to concerns about the robustness of the standard set at level six.
54. The meeting decided that fairness to pupils was the paramount issue. However, it will be important that, in future years, tests contain enough material to make sure that level six awards can be made on the same basis as those for other levels.
55. RM had produced a detailed pre-test report, which informed the release of the summative test.

56. RM will produce further reports on the awarding process, and the validity of the summative tests. These will aid understanding of the tests' validity in the final evaluation report.

4.1.2 Evaluation of objective one

57. This interim report will make a judgement as to whether the 2005 pilot has achieved each of its objectives. It will make a judgement in each case if there is enough available and clear-cut evidence.
58. Of all the objectives, validity is the hardest to call. This may reflect the fact that validity is the most important concept in evaluating a test's fitness for purpose. Therefore, the interim report makes no statement as to whether objective one is likely to be successfully achieved in the 2005 final evaluation report.

4.2 Objective two

59. Objective two is:

> Confirm that the infrastructure software (CPS, APS and DPS), and RM processes for supporting schools during the pilot (technical and customer services support facilities) are scalable for use with a full national cohort and perform their functions without failure.

60. This objective has been amended upon DfES feedback to the interim evaluation report product description. Thus, it refers not only to technical issues relating to the infrastructure software, but also to processes for supporting schools.

61. The CSF associated with objective two is:

> Infrastructure scalability and reliability
>
> This CSF is met if the infrastructure software supports the connection of all pilot schools with CPS availability of 99.5% or greater for all schools.

62. This CSF was written before the amendment to objective two described in paragraph 60. Therefore, it only refers to the first half of the amended objective.

63. Other Success Factors associated with objective two are:

- Infrastructure software has no pre-test evidence of critical faults when released to pilot schools.
- 95%+ of the functionality used by schools within the APS works.
- Majority of schools report that infrastructure software is straightforward to install and performs its functions well.

4.2.1 Findings

4.2.1.1 Software specification and testing

64. The major areas of development for 2005 were derived from the 2004 lessons learned report (see paragraph 198 below).

65. A problem identified in 2004 was that different branches of the consortium were using separate fault-logging systems. This had led to confusion and some faults being overlooked. For 2005 this problem was remedied by having the whole consortium move over to the RM fault-tracking database, known as Test Director.

66. There was evidence that, on transfer to Test Director, some 'Issues Raised' (IRs – or faults) had been inappropriately closed, but not fixed. These were noted by QCA; 3T were required to 'tidy up the database' at a later date, so that a full audit trail of activity could be maintained.

67. In 2004, it had also been felt that RM's reach into its subsidiary 3T had not been sufficient and that this had led to the head contractor not being aware of some

problems early enough. This was remedied in 2005 by the appointment of an RM employee to manage 3T's testing.

68. Many and varied types of software testing were undertaken. These included:
    - ongoing quality assurance checking
    - due diligence checking (by RM of its subsidiaries, and by QCA of RM)
    - acceptance testing – once again, by both RM and QCA

69. Mike Peppiatt, the QCA technical authority, was involved in all aspects of software testing. He did not feel that his involvement in early quality assurance compromised his later participation in acceptance testing.
70. There were far fewer errors in software coding than was necessary to meet industry standards. This was so, even though the infrastructure software was, by 2005, an exceptionally large development.
71. Where individual errors did persist, there was a perception that it was taking a disproportionate amount of time to fix them. This was because all widely occurring errors had been noticed and fixed, and the remainder were complex errors that only occurred intermittently in obscure, difficult-to-reproduce circumstances.

4.2.1.2 Installation of infrastructure software

72. Evidence presented by RM on releasing the infrastructure suggested that installation of the APS took between one and three hours. Installation of the DPS was reported to take approximately 30 minutes. This was felt to be a reasonable amount of time by school staff who were asked.
73. A small number of schools (nine) responded to a questionnaire item about APS and DPS installation. In both cases, this small sample of school staff returned results that showed them to be highly satisfied with the process.
74. Two due diligence school visits by a QCA researcher backed up RM claims that school staff found the installation of the infrastructure to be satisfactory.

4.2.1.3 Summative pilot

75. The release recommendation for the summative pilot materials listed eight defects with priorities of 'urgent' or 'very high'. QCA's technical authority felt that all these defects would occur in relatively rare circumstances, had been extensively tested, and there were legitimate reasons why fixes could not be provided in time for live summative release (e.g. because the fault was difficult to replicate in the testing lab). He contrasted this state of affairs with that existing

before the 2004 pilot; this year far fewer errors were present and a much more robust software solution was sent out to schools.

76. Figures presented to QCA by RM showed that 45,595 unique, complete pupil data files had been returned to the central server. These files came from 402 schools.
77. Thus, the number of pupil data files was well over the minimum acceptable number of 12,000. This was an important finding to demonstrate the technical scalability of the software solution (and would also permit robust analysis of validity).
78. Whilst a formal report has not yet been made to this effect, the QCA technical authority believed that the CPS server had been available for all schools throughout the entire summative pilot window.

4.2.1.4 Technical and customer services support facilities

79. Further information on customer services and technical support will be provided in the final evaluation report.

4.2.2 Evaluation of objective two

80. The Key Stage 3 ICT software was much more robust and reliable in 2005 than had been the case in 2004. A transparent and defensible software testing regime was implemented, leading to a product which – although it did have some pre-release defects – was able to return over 45,000 pupil data files. Further, the central server, which co-ordinated the whole operation, was believed to have remained available for the duration of the pilot.
81. For these reasons, the pilot has achieved the technical scalability and reliability aspects of objective two.
82. In contrast, there is relatively little information upon which to judge the scalability of support services. The final evaluation will report information that becomes available and make a judgement with respect to this facet of the objective.

4.3 Objective three

83. Objective three is:

> Provide all schools participating in the 2005 pilot with accurate formative reports from the practice test and an accurate summative report from the summative tests.

84. The CSF associated with objective three is:

> Accurate formative and summative reports
>
> This CSF is met if the formative and summative reports produced accurately reflect the activities undertaken by pupils and testers and the majority of schools report finding the reports useful.

85. Other Success Factors associated with objective three are:

- Schools' feedback confirms the formative and summative reports are useful and in a user-friendly format
- Statements in summative reports are perceived by schools to be consistent with the NC levels awarded by the test
- The automated marking is generating statements for reports that accurately reflect what pupils have done

4.3.1 Reporting functionality

86. Formative reports were developed to be delivered on completion of practice tests. The intention was that these reports would be available on screen, or to print out for distribution to pupils. It was also intended that these reports would be available to teachers via the APS.
87. Formative reports gave pupils a brief summary of the tasks that they had undertaken in the test, and feedback in respect of the nine capabilities of ICT. These capabilities have been developed by QCA, and amount to a sub-division of the National Curriculum Programme of Study.
88. Summative reports were designed to be based upon pupils' performance in the summative test. They have not yet been delivered to schools as this report is written.

4.3.1.1 Pre-release testing

89. RM investigated the plausibility of reports as part of their acceptance testing prior to product release. RM staff conducted test sessions and then printed out the resulting reports and made sure that what they saw was plausible and realistic.

4.3.2 Opinions about formative reports

90. Teachers were asked how accurate the formative reports at the end of the practice test were. An initial collation and analysis of the responses to that questionnaire item showed that approval for the formative reports was extremely low. The average rating on a scale of 0 – 10 was 3.13.
91. Twenty-eight teachers also wrote longer comments in response to this item in the questionnaire. A large majority of these comments (20) indicated that teachers and pupils had either not received the formative reports, or had not used them.
92. The reasons for teachers not finding or not using the formative reports will need further probing. Several teachers said that they had not seen formative reports at all (e.g. *Didn't get one!?*, *What statements?*). This may imply that, in some schools, formative reports were not being produced from the practice test due to an intermittent software defect. (RM report an open service call to investigate this issue.)
93. However, other comments suggest different reasons for teachers and pupils not taking advantage of formative reports. Examples of this include:
    - *Very strict timetable so didn't investigate the report at the end of the practice. We had lots of machines crashing*.
    - *Did not use this feature as we did not know these tests could be scheduled – thought it was for the proper tests only*.
    - *I didn't see it, I was too busy observing the 'shut down' process.*

94. There was further comment on formative reports from teachers who did see the reports. Most of it was negative:
    - *Did not relate to what the pupils did during the practice test.*
    - *A little too bland. If the students did not match the task timeline it meant they got a poor comment and it was very very difficult to work out what was required.*
    - *It was impossible to make any sense of it, as we had no idea what they had done right or wrong …*

95. Teachers' negativity towards the formative reports related both to their usefulness and their accuracy.
96. Several teachers also commented that they would have liked to see estimated levels delivered from the formative report.
97. A group of teachers at a QCA stakeholder group gave feedback that corroborates the findings from the RM questionnaire. Most teachers had not used the formative reports, and there was a lack of enthusiasm for them amongst those who had.

98. The problems in the area of formative reports could come from several sources. Possible causes might include:

    - There is an intermittent software defect that prevents some schools from seeing formative reports.
    - Practical issues (such as pupils rushing to log out of workstations at the end of sessions, and so not viewing the reports, or school staff not being aware of all the functions in the APS) prevent schools from taking full benefit from formative reporting functionality that does actually exist.
    - ICT teachers are not particularly good at doing formative assessment.

99. None of these possibilities can be completely proven or ruled out at the time of writing. The implications of each potential cause of the problem are very different. Diagnosis of the problem must not be limited to one approach (for example, the project should not merely look for a software defect, and then assume that the wider issue of formative reporting is resolved, if the defect is fixed, or if none is found).

100. The QCA has recognised that the formative aspect of the project has not been prioritised to date, and has escalated formative reporting to the issues log of its Content and Mark Schemes Working Group.

4.3.3 Evaluation of objective three

101. An evaluation will be made on the summative reports in the September evaluation report.

102. The very low approval rating for the formative reports, and the widespread negativity amongst those who commented on them, means that it will be very difficult for the 2005 pilot to succeed against this aspect of objective three in the final evaluation report.

4.4 Objective four

103. Objective four is:

> Carry out an investigation that will test whether a scalable (national cohort) system achieves the desired test security in relation to: data randomisation, the test window, and security breaches and hacking.

104. The wording of this objective was changed, following agreement of the project board and the OGC gate 4(b) review team. This re-wording inserted the notion of an 'investigation' of test security, rather than requiring an evaluation of whether the test solution had been fully secure in 2005. The effect of this change in wording was to make the objective easier to achieve. This issue is discussed further in paragraphs 138 to 140.

105. The CSF associated with objective four is:

> Test security
>
> This CSF is met if the test and test data is handled by the system in a secure manner with test data returned securely to the CPS and results returned securely to schools. The 2005 trust management[1] report commissioned by QCA will help inform whether this CSF has been met. The validity report will also provide evidence that the test was secure.

106. This CSF was written before its governing objective was amended, and so does not reflect the change in the wording of objective four.

107. Other Success Factors associated with objective four are:

- evidence from RM's security audit log that system security was not breached
- feedback from schools confirms that pupils are unable to cheat by looking at other pupils' PC screens

4.4.1 Nature of security issues

4.4.1.1 Classroom issues

108. The aim of the Key Stage 3 ICT tests project is to develop a system that will deliver high-stakes National Curriculum tests securely. As a step along the road to secure, high-stakes delivery in 2008, the 2005 pilot was run as if it was high-stakes (see paragraph 187).

[1] Trust management has been defined as follows: 'Trust management is concerned with ensuring that all storage and electronic movements of confidential project materials between different parts of the test system is done as securely as is needed. It is about the ability to transmit, collect, store and process information electronically and to ensure the confidentiality, integrity and availability of the KS3 ICT System at all times.'

109. A major potential source of insecurity in the testing system arises from the nature of ICT rooms in English schools. In such classrooms, pupils' computers are generally very close together (pupils can be as little as 12 – 18 inches apart). Thus, there has been concern that pupils might copy from each others' screens during the tests.

110. ICT rooms are typically fully booked during the school term (for instance, they are used by many curriculum subjects, not just ICT). As such, it can be hard for schools to schedule summative test sessions for whole cohorts of Year 9 pupils. In order to facilitate the scheduling of test sessions, the project developed the approach of running the test in a four-week window (see paragraph 191 below). However, the security concern arising from a relatively long test window was that test content might become known (to pupils or teachers) and then be improperly communicated – e.g. to subsequent classes in a school, or to fellow pupils or teachers in another school.

111. The test development project found different solutions to the potential threats to security described in the preceding two paragraphs. To counter the possibility of pupils copying from very close neighbouring screens, there was a policy of providing multiple (cloned) test versions and randomised surface data in assets and task instructions (see paragraph 214).

112. No specific tactic was implemented in the 2005 pilot to counter the threat of test content becoming known in the early part of the test window, and then being improperly communicated to third parties. The project is discussing strategies for the 2006 test to counter any such security breaches. The current front-runner is a proposal to publish a description of test content to all schools before the test window. It is surmised that this publication will remove any advantage that might be gained from communicating test content during the window.

113. The viability of this potential approach is currently being evaluated by the project, and no final decision has been made to implement it in 2006.

4.4.1.2 Institutional issues

114. The deployment of the Key Stage 3 ICT tests in classrooms gives rise to some security concerns. However, security of information (either confidential test content or sensitive data on persons or institutions) could be compromised at several points during the test development and delivery cycles.

115. The potential for 'institutional security breaches' applies both to physical loss of sensitive information (e.g. if a room containing drafts of test content were

insecure, or if a briefcase were left on a train) and to electronic loss (e.g. as a result of hacking).

4.4.1.3 Available information

116. The amount and quality of information on the two types of security issues described in the previous sub-section is starkly different. Relatively little information is yet available on potential classroom breaches of security, whilst a large amount of work has been done in the following areas:

- RM's infrastructure for delivering the tests
- Methods of encryption to ensure that unauthorised third parties cannot tamper with test or results data
- Procedures and policies to back up security

117. Such potential breaches of security are grouped under the sub-heading of 'institutional issues'.

118. RM are currently conducting an experiment in which pupils are actively encouraged to copy. The pupils and their teachers will then be asked whether they thought copying was successful. This work will be summarised in the final evaluation report.

119. Some information as to pupils' propensity to (and success in) cheating can be gleaned from a report of research based on QCA's own visits to schools during the summative pre-test, and from other write-ups of visits conducted as part of the project team's monitoring work.

120. RM's validity report – once again delivered after this interim evaluation report has been written – will include a summary of the security audit log, which will indicate whether the system has been hacked during the summative pilot.

121. The independent trust management specialists, KPMG, have conducted a substantial project on behalf of QCA. This work has reviewed the security systems, processes and policies in place for the tests.

122. The first phase of trust management work was carried out with respect to the 2004 pilot. This work was:

- visits to the RM sites and interviews with key staff to review the set up of the IT hardware and software on which the Key Stage 3 ICT System is hosted (technically called a Web Hosting Infrastructure Review)
- an external review of how effectively the Key Stage 3 ICT System stood up to a defined set of external 'attacks' (technically called Penetration Testing) followed up by on-site investigation.

123. Reports from these activities provided a set of action points, which have either been implemented by RM for the 2005 pilot, or have been deferred for 2006 development.

124. With respect to the 2005 pilot, two pieces of work have been undertaken. Firstly, the Web Hosting Infrastructure and Penetration Testing investigations were repeated by KPMG on the 2005 Key Stage 3 ICT System before the delivery of the 2005 pilot in April/May. Also, KPMG reviewed security processes and procedures in the project and the security of the test development and management process.

125. In addition to these formal reports, there has been a programme of meetings relating to trust management. These had various purposes, and different meetings included KPMG, QCA and RM. One set of meetings was attended by QCA and KPMG, but not RM. These meetings were for KPMG to hand over reports to QCA, and discuss their implications frankly. In this report, reference is made to the minutes of the 23rd May 2005 report handover meeting.

126. The Office of Government Commerce's Gateway Review (see page 33, below) made several recommendations and general comments with respect to security and trust management.

127. The QCA project team has presented a paper to the programme board on related topics. The paper contained a number of recommendations and proposals for further action.

4.4.2 Findings

128. QCA researchers visiting pre-test schools noted whether tests were run under 'exam conditions' or not. They considered that three test rooms were run under exam conditions, whilst six were not.

129. The reasons for exam conditions not pertaining included:

- Technical problems causing pupils to lose concentration and then to 'take the test less seriously'.
- Teachers not really following through on conducting sessions under exam conditions.
- Teachers not understanding the significance of a summative pre-test – i.e. that it contained the same material as the full summative test, and was not just a practice session.

130. QCA staff visited schools in the summative window, in addition to the pre-test visits. Initial impressions were that the issues of exam conditions and security of test content were being taken more seriously during the summative window.

131. KPMG's 2005 penetration testing demonstrated that lessons had been learnt from the Phase 1 testing undertaken in 2004. KPMG's view was that there were far fewer vulnerabilities exposing the KS3 ICT system to risk in 2005 than there had been in 2004.

132. The review of the 2005 infrastructure (Web Hosting Infrastructure Review) also demonstrated that lessons had been learnt from 2004. Many of the Phase 1 report recommendations had been taken up and addressed or upgrades to the system in autumn 2005 were planned.

133. The new KPMG reviews conducted in 2005 produced less satisfactory findings. The reports on security policies and procedures, and test development and management suggested that substantial work was needed to ensure the security and confidentiality of commercially or personally confidential information.

134. KPMG commented at a meeting that there was substantial work to ensure the security and confidentiality of commercially or personal confidential information across the RM Consortium, but special concern was reserved for 3T, where a substantial number of vulnerabilities had been observed.

135. KPMG were asked to provide a note of any additional concerns to Martin Ripley, the project's Senior Responsible Officer (SRO). KPMG has not, to date, documented any additional concerns. However, QCA are already implementing a more pro-active management of RM's security policies and procedures, where these impact on the Key Stage 3 ICT tests.

136. The OGC report made several observations with respect to trust management:

- Trust management work had not yet extended to schools or third-party managed service providers.
- Plans for responding to any deficiencies were already in place.
- The fact that these plans were already in place, the high level of diligence on this matter in the project and the relatively limited scope of the 2005 pilot meant that the unproven nature of the trust management regime in the solution should not delay the 2005 pilot.

137. QCA has put in place a programme of work to respond to these observations. Some of the responses to OGC recommendations were in QCA's paper to the programme board, which made the following proposals:

- QCA should conduct a small-scale study to investigate security vulnerabilities in schools. (The small-scale study might raise recommendations for the test development project, and, more widely, for Becta and the DfES.)
- Becta should commission a wide-ranging study into security vulnerabilities to report in summer 2006.
- The provision of data to the DfES by the QCA should be prioritised in the latter's programme of work to comply with British Standard (BS) 7799[2].
- The proposals should apply to service providers, as well as schools.

[2] BS 7799 is a standard which sets out criteria and best practice for an organisation seeking to protect information held on IT systems – it is concerned with policies on passwords, access rights and how change is approved when it impacts on IT security, for example.

4.4.3 Evaluation of objective four

138. The wording of objective four was changed – so that, rather than requiring a test of whether a scalable system actually achieved the necessary security, the objective would be met if it could be demonstrated that a credible investigation had been conducted into this issue.

139. The KPMG work has been carried out. Lessons were learned from the 2004 reports, and there is reason to believe that 2005 reports will be addressed with equal diligence. Therefore, the part of the objective dealing with 'institutional security issues' has been achieved.

140. RM's experiment to see whether pupils can successfully copy from neighbouring screens will be summarised in the final evaluation report. This document will then make a final judgement on the 'schools issues' part of this objective.

141. The approach taken to trust management in 2005 can be described as follows:

- The issue of information security in a high-stakes e-assessment is a new question for the QCA.
- The correct level of information security for an on-screen National Curriculum test is not an established concept.
- Security has been actively considered throughout the development of the KS3 ICT tests.

142. This approach amounts to an 'investigation' of the security of the system (as defined in the re-worded objective, see paragraph 104). In 2006, a more comprehensive test of a fully operational secure system will be conducted. This will amount to an 'evaluation' of the system.

143. The project team believes that this is an informed and appropriate way to manage information security.

4.5 Objective five

144. Objective five is:

Ensure that schools that have volunteered for the pilot and meet the minimum specification have a satisfactory experience, even if they are unable to participate in the April/May test window.

145. The CSF associated with objective five is:

School experience

This CSF is met if the schools who meet the minimum specification and complete technical accreditation report that they had a satisfactory experience, with average customer satisfaction reported by schools of at least 7.0 out of 10.

146. Other Success Factors associated with objective five are:

- All accredited schools not participating in the April/May test window are able to run the summative test before the end of the school year.
- Majority of schools report satisfactory experience.
- Positive feedback from schools on contact with RM, and quality and helpfulness of materials.
- Positive feedback from schools about manageability of test requirements.

4.5.1 Available information

147. This section of the report is based on the following sources of information:

- Analysis of questionnaires on the familiarisation materials, practice tests, and summative pre-test
- Write-ups of QCA project staff's monitoring visits to schools
- An evaluation report based on visits to schools made by QCA research and ICT curriculum staff during the pre-test window
- Notes of QCA Teacher User Groups.

148. The report does not consider the following types of information:

- Write-ups of RM staff's visits to schools that were appended to the TDA's release recommendations.
- Spreadsheets containing 'raw data' from summative pilot questionnaires that had been made available to QCA by RM.

149. The reasons for not using these sources of information were, respectively:

- RM staff's writing up of school visits was very variable.
- The summative pilot questionnaire data were, as yet, unchecked and unanalysed.

150. An important aspect of this objective appears to be manageability. Manageability has been addressed tangentially in various places in the project, but has not been systematically and thoroughly studied by the project.

151. Indeed, the QCA project team has suggested that a Regulatory Impact Assessment (RIA) should be carried out before the Key Stage 3 ICT tests are identified as a statutory requirement for schools. The responsibility for such an impact assessment would rest with the programme, not with the test development project. (See page 34, below, for a description of the Key Stage 3 ICT assessment programme).

152. A Success Factor governed by this objective relates to the experience of schools that were not able to take part in the April/May 2005 pilot. It is not possible to report their experience 'before the end of the school year', since the school year has not yet ended at the time of writing.

4.5.2 Findings

153. The findings with respect to this objective are organised according to major phases of the pilot.

4.5.2.1 School accreditation

154. QCA has expressed some concern that, whilst over 2500 schools expressed an interest in taking part in the 2005 pilot, only about 400 ended up sending valid data from the summative pilot. A particular concern was that the accreditation process might have been too slow, or otherwise onerous, for schools. However, when teachers in a QCA feedback group were asked this, they felt that the accreditation process had not been overly slow or onerous.

155. The 402 schools that returned data represented about ten per cent of secondary schools in England. The TDA has produced evidence that these schools are representative of the country's schools with respect to several background variables (e.g. school size, type, geographical location, etc.). However, convincing evidence has not been produced as to the quality of ICT provision in pilot schools. A concern would be if the 2005 pilot schools were the ten per cent that were best at teaching ICT, had the most up-to-date equipment, had the best attainment, and so on.

4.5.2.2 Familiarisation materials

156. The Test Development Agency's pre-test report contained findings from familiarisation, practice test, and summative pre-test questionnaires. There was a questionnaire for teachers and one for pupils, in each case.

157. The familiarisation questionnaire had several positive findings. Approximately 80 per cent of teachers who responded agreed, or strongly agreed, that the

familiarisation materials were useful for pupils. A similar percentage of teachers agreed, or strongly agreed, that pupils had had enough time to complete the familiarisation sessions. Another positive finding was in the reaction to toolkit applications. All were agreed to be useful by large majorities. (The smallest figure for those agreeing, or strongly agreeing, that an application was useful was the database – which had a 66 per cent agreement rate.)

158. There were some indications of positive pupil opinion towards the familiarisation materials. Slightly over half of a sample of over 600 pupils found familiarisation materials interesting, whilst nearly 80 per cent found these materials useful. As in the case of teachers, clear majorities of pupil respondents found the applications easy to use. The least usable was the database, which 57 per cent either agreed, or strongly agreed, to be easy to use.

159. There were some results that cast the benefits of the familiarisation materials in a more doubtful light. Teachers were asked whether 2 x 15 minute sessions were enough for pupils to familiarise themselves with the toolkit. There was only approximately 50 per cent agreement with this proposition.

4.5.2.3 Practice test

160. Some teacher questionnaire items related to the practice test. A majority (56 per cent of a sample of 47 respondents) thought that the practice test allowed pupils to demonstrate their ICT capabilities.

161. Pupils' feedback in respect of the practice test showed that approximately half found it interesting, but only 45 per cent had enough time to finish the practice tests. However, a larger percentage (58 per cent) agreed, or strongly agreed, that the practice tests allowed them to demonstrate their ICT abilities.

4.5.2.4 Summative pre-test

162. Summative pre-test questionnaire results largely backed up points made in respect of the familiarisation and practice sessions. A small sample of teachers gave their views, but they were largely positive in respect of:

- pupils finding it useful to do a practice session before doing the summative pre-test.
- pupils finding the pre-test interesting.
- pupils being able to demonstrate their ICT capability in the pre-test.
- pupils having sufficient time.

163. Once again, there were strong majorities in favour of the usefulness of toolkit applications. The exception was the database, which only half of this small sample believed to be easy to use.

164. Pupil pre-test questionnaires provided positive findings in respect of:

- the usefulness of sitting a practice test prior to doing the pre-test
- being able to demonstrate ICT knowledge and abilities
- the ease of use of all applications (including the database)

165. However, pre-test questionnaire findings showed cause for concern, since a majority of pupils gave negative responses in the following areas:

- the test being interesting
- having enough time to complete the test

166. An evaluation report was written by a QCA researcher, based on visits to schools during the summative pre-test window. Important emerging issues included:

- Schools finding it hard to fit 2 x 50 minute sessions into timetables, and/or
- Fitting the sessions into timetables being disruptive of other lessons – either if pupils arrived late to the lesson after an ICT test had overrun, or if other subjects were not able to use an ICT suite during the pre-test period.
- There were variable findings as to the number of members of staff that were needed to invigilate test sessions. There was some evidence, from several different schools, that more than one teacher and an ICT technician were being used to invigilate. However, it was not always clear whether every member of staff present in test rooms was absolutely essential, or whether some were there 'out of curiosity'.

4.5.2.5 Summative pilot

167. Full findings in regard of schools' experience in the summative pilot are not yet available. Further findings will be reported in the final evaluation report in September 2005. However, some emerging issues can be noted.

168. Some schools have commented that, although the test software seemed overall to be quite robust, there tended to be one or two pupils in each session whose tests crashed. This could be a problem with multiple causes, but which could give rise to several problems. In terms of causes, in the case of odd machines in a session crashing, there were strong grounds for believing that the problem was local, rather than an issue with the QCA system (e.g. a particular workstation had poor specifications, or was at the end of a network).

169. However, when one or two machines in a session crashed, several problems could result. These are listed below:

- Schools trying to run session one and session two 'back-to-back' could find session two delayed, due to just one machine having problems. This could delay the second session for a whole class.
- It could be perceived as unfair if one or two pupils engaged with the test, but, due to technical problems, were not awarded a level.

- Schools perceived pressure to make sure that all pupils – even those who had been absent for their scheduled session – completed tests during the window. This problem was made worse if some were seen to 'be given a second chance' due to technical problems.
- One teacher reported running special 'catch-up' sessions for such pupils. He found this to be a clumsy work-round solution, however.

170. A QCA Teacher User Group was concerned that the tests generated significant amounts of administration. A consensus emerged in this group that Exams Officers in schools should, in future, play a significant role in administering the Key Stage 3 ICT tests. However, the group's view was that, at present, many Exams Officers do not have the necessary IT skills to perform this function adequately.

4.5.3 Evaluation of objective five

171. There was a weight of opinion from teachers (and to a slightly lesser extent from pupils) showing approval for these tests. There were also significant concerns about the tests' impacts, as well as important findings that remain to be reported in the final evaluation report. However, the balance is that the range of positive opinions shows that objective five has been passed; albeit with concerns about the changes that will have to take place in schools to make the running of these tests straightforward.

4.6 Critical Success Factor six

172. CSF six relates to the Office of Government Commerce Gateway review 4b. It does not pertain directly to any one objective.

173. CSF six is:

> Following on from the OGC Gateway review of the two-per-cent Technical Pilot (Gateway review 4a) the OGC will review preparedness for a national rollout of the test infrastructure software and accompanying familiarisation materials.
>
> This CSF will be met if the OGC 4b review is green.
>
> If the 4b OGC report is amber the CSF can only be met if the recommendations can be implemented successfully, allowing the project to proceed.
>
> This CSF will not be met if the OGC review 4b is red.

4.6.1 Findings

174. The primary purposes of an OGC Gateway Review were to confirm that

- contractual arrangements were up to date.
- necessary testing had been done to the client's satisfaction.
- the client was ready to approve implementation.

175. A Gateway 4b review was a 'Ready for Service' assessment of the infrastructure and associated familiarisation materials.

176. The OGC review was conducted in March 2005.

177. The overall status of the project was found to be amber. The efforts and capabilities of the QCA project team were warmly praised in the report. Particular examples of good practice were said to be:

- the scale and depth of stakeholder engagement
- the pro-active diligence of QCA project management
- the co-ordination and management of the whole (software) testing regime through RM's database, Test Director

178. It is believed that the OGC's recommendations can be managed via the OGC issues logs that are regularly presented to the project and programme boards.

4.6.2 Evaluation of Critical Success Factor six

179. Critical Success Factor six has been achieved.

5 Annex A: Background

5.1 The Key Stage 3 ICT assessment programme

180. The Key Stage 3 ICT assessment programme has three strands:

- Test development
- Preparing schools/the ICT strand of the Secondary Strategy
- Infrastructure in schools

181. Whilst the test development strand of the programme must develop a reliable and valid test, this cannot in itself ensure that schools effectively take up the new test, and that the development will be, in the widest sense, successful. The responsibility of the other strands – for example, to make sure that teaching is appropriate, or that infrastructure in schools meets minimum specifications, will affect the success or otherwise of the test development project.

182. The Key Stage 3 ICT assessment programme is managed by a programme board. This board is chaired by Andrew McCully of the DfES. Correspondingly, the test development project has a board. Martin Ripley of the QCA is chair of the project board, as well as the project's Senior Responsible Officer (SRO).

5.2 The Test Development Agency for the KS 3 ICT tests

183. The Test Development Agency (TDA) is a consortium led by Research Machines PLC (RM); a provider of ICT software, services and infrastructure to UK educational institutions. RM is the prime contractor responsible for overall project delivery. The consortium also includes RM's subsidiary, 3T, who develop the test software, including the ICT toolkit, and Tata Infotech, who are responsible for infrastructure software.

184. For the 2005 pilot, the measurement expertise in the project was strengthened by the appointment to the consortium of the Centre for Formative Assessment Studies (CfAS) from the University of Manchester. In particular, the experienced and respected analyst, Nick Nelson, of CfAS, has taken a major role in the analysis of 2005 pilot data.

185. RM's internal expertise in assessment has been strengthened by the appointment of Miranda Simond as Assessment Manager.

5.3 Pilots in the Key Stage 3 ICT test development project

186. A technical pilot of the Key Stage 3 ICT tests was carried out in 2004. An evaluation report describing that pilot was sent to the DfES on 19th November 2004.

187. The 2005 pilot has been characterised as the first step on the way to a full national rollout. The 2005 test was not considered a high-stakes test in and of itself, but it was treated as a pilot for a test of such stakes; thus, procedures and guidance for administering the test were similar to those that would be in place in a high-stakes administration.

188. The 2005 pilot aimed to get between 500 and 600 schools to participate in the summative pilot. It was envisaged that this number of schools would lead to a data set of at least 12,000 pupils.

189. There was a considerable amount of activity in the 2004/05 school year. Firstly, schools had to pass the accreditation process and have their IT systems 'health checked' (i.e. checked to make sure that they were of a minimum technical specification), in order to take part in the pilot. Then, they had to install the infrastructure software from CD-ROM. Next, schools downloaded familiarisation materials via the infrastructure software. These sessions allowed pupils and teachers to get acquainted with the test model, and (in the case of school staff) aspects of the administration process. There were also two fifty-minute practice test sessions available to schools. These were based on the 2004 test.

190. A small number of schools took part in a summative pre-test in March 2005. It was intended that this pre-test would collect data from 500 pupils for each test form. Analysis of these data would inform the release recommendation for the full summative pilot.

191. The summative test pilot was run over a four-week window (25th April – 20th May 2005). This was to allow schools to schedule test sessions for the whole of their Year 9 cohorts – for example, taking into account situations where the availability of ICT rooms was limited due to their use for other learning activities in the school.

192. The DfES White Paper on 14-19 Education and Skills refers to the Key Stage 3 ICT tests as follows:

> We are developing a test in ICT to build on the existing practice of teacher assessment. This will be an online *(sic)* assessment and will be electronically marked. It will be introduced alongside the other external tests at age 14 from 2008, subject to a successful pilot.

193. In order to deliver a statutory ICT test by 2008, the project intends to conduct pilots in subsequent years as follows:

2006: National pilot
2007: Full dry run, 'as statutory'

5.4 Elements of the Key Stage 3 ICT test delivery system

194. The main Key Stage 3 ICT test delivery system consisted of the following major elements:

- a Central Point System (CPS); a central administration database hosted by RM at Telehouse (London)
- an Administration Point System (APS), sitting on the server in each school
- several Delivery Point Systems (DPSes), sitting on workstations in each school

195. The delivery system was responsible for the secure transmission, storage and management of the following elements:

- a set of test packages that collectively constituted a single ICT test
- a test results package for each test taken

196. The CPS could also be used to distribute software upgrades for either the APS or DPS.

197. The majority of the administration system had been developed before the 2004 pilot of the KS3 tests. In 2005 major areas of development implemented recommendations from the 2004 lessons learned report.

198. That lessons learned report emphasised that the reliability and usability of the infrastructure software must be improved. The following areas were highlighted:

- improved reliability of test recovery
- proxy server authentication
- improved clarity/ease of use within scheduling activities
- improved test recovery (i.e. not manually from each station that had lost contact with the server)
- workstation deployment (use of a package that installed workstation software via a central server, not manually to each machine), to reduce the potential burden upon schools.

199. In addition to these developments to the administration system, two new applications were developed for the application toolbox: database and presentation software.

200. The software that was released to schools carried the QCA logo. Thus, whilst the project is still in pilot phase and QCA has not yet accepted any finished product, schools tended to refer to 'the QCA software'. This fact is reflected in the way that the software is referred to in this report.

5.5 A description of the tests used in the 2005 pilot

201. The Key Stage 3 ICT tests addressed the construct of ICT capability, which was defined as follows:

> 'ICT capability is about having the technical and cognitive proficiency to access, use and communicate information using technological tools.
>
> Learners demonstrate this capability by purposefully applying technology to solve problems, analyse information, develop ideas, create models and exchange information.
>
> They are discriminating in their use of information and ICT tools.'

202. ICT capability is often contrasted with ICT skills, which are the technical competences necessary to do simple tasks using commonly-used software applications.

203. The test reported levels in terms of National Curriculum (NC) level descriptions. A key aid in allowing the NC levels to be operationalised in an e-assessment was the QCA Rules Base. The Rules Base is a sophisticated branching database in which level descriptions are broken down into separate granularities of evidence (known as process indicators and elaborations).

204. A small extract from the Rules Base is shown in the figure below.

Level Description sub-division	**Granularities of the Rules Base**	
	Process indicators	**Elaborations**
(A) Pupils select the information they need for different purposes, check its accuracy and organise it in a form suitable for processing.	i. (A) Select information/assets for specific purposes	(b) Check accuracy by finding information/assets from more than one source (i, ii)
		(c) Check validity by finding information/assets from more than one source (i, ii)
		(d) Select relevant parts of the information/assets gathered, ignoring irrelevant parts (i, ii, iii)
	ii. (A) Organise information/assets for processing	(r) With guidance select technology tools for problem solving and decision making (i, iii, iv)
		(y) Select and apply technology tools for information analysis (ii)

Figure 2: A schematic diagram of a part of the Rules Base[3]

205. Figure 2 shows a particular part of a level description (A), which is sub-divided into two processes (i and ii). Each of these processes maps to one or

[3] This diagram originated in an RM report.

more elaborations. For example, process indicator i is mapped to four elaborations in this diagram (b, c, d and r). Equally, each elaboration can be mapped to several process indicators (e.g., elaboration b is mapped to process indicators i and ii).

206. In this report Rules Base layers along the horizontal axis of Figure 1 are referred to as 'granularities', because process indicators relate to relatively coarsely-grained evidence, and elaborations finely-grained evidence.

207. RM has developed further granularities, in order to facilitate the use of the Rules Base in an e-assessment. Captured actions are essentially keystrokes, mouse clicks, and so on – pupils' actions that can be captured by the computer during the test. Opportunities are a downward extension of the QCA Rules Base, developed by RM. They are a middle step between elaborations – the most finely-grained part of the Rules Base, and captured actions.

208. Opportunities can also be described as sequences of actions that pupils could carry out in a test. Opportunities have a hierarchical relationship with elaborations; either one opportunity can make up an entire elaboration, or an elaboration can be the combination of several opportunities. There has been a debate within the project about the relative merits of elaborations and opportunities. In simple terms, RM has stated the view that opportunities are the lowest granularity that can provide meaningful evidence of ICT capability; QCA has taken the line that elaborations are the finest grain at which ICT capability can be described.

209. The tests used for the 2005 pilot reported NC levels between three and six[4]. The testing model was based on two tiers; pupils who were given an Initial Level Assessment (ILA) of three or four were entered for the level 3 – 5 test, and those with an ILA of five, six or greater were entered for the 4 – 6 tier.

210. Each test consisted of five tasks, split over two 50-minute sessions. Session one consisted of three tasks of roughly equal length, and session two had a 17-minute-long task, followed by a 33-minute task.

[4] RM are currently conducting a research project that will lead to a recommendation as to the best way to assess NC levels seven and eight. However, this work is not part of the 2005 pilot, and so will be reported in neither the interim nor final evaluation reports. The QCA will present its recommendations as to how to assess levels seven and eight to the DfES in a separate document.

211. All tasks were distinct from each other and self-contained. Tasks were also based on different contexts; there were four contexts across the five tasks (i.e. two tasks in each test had the same context).

212. In general, task 1 in any test form was the easiest task in the test, and tasks got more difficult throughout the test until task 5 – the most difficult.

213. This was a linear test in that pupils were presented with tasks in an order that was determined before the start of their tests, and which did not depend upon the nature of their responses. This was not an adaptive test.

214. There were two forms of each tier of the test, making four forms in total. The forms were labelled as A or B. In each case, form B was a clone of form A. A task was cloned in that it was essentially the same as its parent, but was given some surface-level change (for example it was set in a different context). The use of cloned forms was intended to make it more difficult for pupils at neighbouring computers to copy from each other.

215. Data within pupils' tests were randomised. Randomisation meant that surface information (e.g. numbers, names, etc.) in assets (documents, spreadsheets, etc.) or task instructions were varied.

216. Randomisation is carried out on a per task basis. Each task had four sets of data attached to it. There are five tasks in each test and so the number of different variants of one test was 4x4x4x4x4; that is 1024. Further, there were four test forms – therefore there were 4096 different variants.

5.6 Evaluation objectives

217. Objectives for the 2005 pilot were established by the DfES through the programme board. The text of each objective is set out at the start of the relevant sub-section of the 'Evaluation of specific objectives' section of this report (pp. 7ff).

218. The objectives used in this report make up the version that was included in the product description for this interim report. These objectives were, in some cases, revised following DfES feedback.

219. Associated with each objective is a small number of Success Factors, including one Critical Success Factor (CSF) for each objective. The Success Factors facilitate evaluation of whether or not the pilot has met its objectives.

220. In addition to the five objectives, there is a sixth CSF, which is not associated with any objective. CSF 6 relates to the Office of Government Commerce (OGC) Gateway review 4b. The text of this CSF is set out at the start of the relevant sub-section this report.

6 Annex B: Acknowledgements

Thanks are due to the following people, who gave help in the areas described. Apologies to anyone who helped me, but whom I have neglected to mention here. Whilst I acknowledge the help I have received, responsibility for any errors remains mine.

Person	Organisation	Help provided
Jim Brant	QCA	Information relating to objective 1.
Mike Peppiatt	QCA	Information relating to objective 2.
Steve Suckling	QCA	Information relating to objective 4.
Sue Walton	QCA	Information relating to objective 5, and general assistance relating to scope of evaluation, etc. Factual checking of draft report.
Hakan Redif	QCA	Information relating to CSF 6 and general help in finding documents.
Gill Williams	QCA	Help in finding documents.
Martin Adams	RM	Construction of a detailed table showing what information was available from RM. Factual checking of draft report.
Mike Wright	RM	Responses relating to the amount of data that had been gathered in the summative pilot.
Colin Robinson	QCA	Vetting of draft report.
Tim Oates	QCA	Vetting of draft report.
Martin Ripley	QCA	Vetting of draft report.
Miranda Simond	RM	Factual checking of draft report.
Alison Matthews	QCA	Comments on draft report.
George Vassiadis	QCA	Literal proof reading of report.