Qualifications and Curriculum Authority

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

A report to the Department for Education and Skills

October 2005

QCA/06/2341

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 21st October 2005. It evaluated the 2005 pilot's success against a set of objectives. This final report sets the evaluation's definitive findings for each pilot objective. This final report supplemented an earlier, interim, report that was sent to the DfES on 12th July 2005. The interim report contains some information that is not in the final report; for example, some objectives could already be judged by 12th July. Also, the interim report contains an appendix describing some key background concepts in the Key Stage 3 ICT test. This appendix is not repeated in the final report.

Andrew Boyle
QCA
02 February 2006

Contents

List of tables

List of figures

1 Executive summary

This is the final evaluation of the 2005 pilot of the Key Stage 3 ICT tests. This final report supplements an interim evaluation report, dated 12th July, that was sent to the Department for Education and Skills (DfES).

The 2005 phase of the project culminated in a summative test pilot in April and May 2005. In that pilot, over 45,500 pupils in 402 schools sent valid test data to a central server. National Curriculum levels in ICT have been determined for those pupils, and returned to teachers.

The overall evaluation is that the 2005 pilot was a success – when judged against the overall objectives of the project. This judgement arises from a view of the likelihood that the Key Stage 3 ICT test will be of suitable high quality to be put on a statutory footing in 2008.

This overall judgement is supported by the project's success in respect of 2005-pilot objectives, and by knowledge of current and near-future work to improve the whole Key Stage 3 ICT tests solution.

Thus, turning first to 2005-pilot objectives: Table 1, below, shows that the pilot has achieved all its objectives except for those relating to reporting, and one aspect of the security objective. Whilst this is not a perfect result, it is a strong result for a project at this stage of development.

Also, the following recent developments contribute to this positive evaluation:

- The prompt initial work that has gone on to implement recommendations from the interim evaluation report.
- The several bespoke research activities that are due to commence to investigate different issues pertaining to the test.
- The current moves to find an appropriate regulatory approach for the Key Stage 3 ICT tests.

The following table summarises the evaluation of the 2005-pilot objectives. The evaluations of whether the pilot achieved its objectives should be seen in the context of a test in pilot phase.

Objective number	Objective focus	Sub-focus	Finding
1	Validity	N/A	achieved
2	Infrastructure software and support processes scalability	Infrastructure software reliability and scalability	achieved (in interim evaluation)
		Support processes scalability	achieved
3	Accurate formative and summative reports	Summative reports	not achieved
		Formative reports	not achieved
4	Test security	Classroom issues	not achieved
		Institutional issues	achieved (in interim evaluation)
5	School experience	N/A	achieved (in interim evaluation)

Table 1: Summary of findings for each objective

Additional findings and recommendations are organised under 'positive outcomes' and 'areas for further work' sub-headings. The findings are listed in the order in which they occur in the main body of the report.

Objective one: Validity

Positive outcomes

- Face validity has been established for the test as it relates to levels three to five.
- Reliability indices based on 2005-pilot data demonstrate that this test is capable of providing measurement that approaches the lower bound of what is considered to be acceptable in high-quality testing.
- There are reasons to believe that reliability indices might increase once the test has bedded down more in the English education system.
- Whilst the rate at which the test classifies pupils into the same level on repeat administration seems rather low, this rate is not necessarily incompatible with classification consistency rates in other high-stakes assessments.
- Findings suggest that this test was fair for both genders.
- Pupils who were entitled to free school meals scored significantly less well on the test than pupils who were not so entitled. However, it is reasonable to conclude that the lower scoring by pupils entitled to free school meals reflected the genuinely lower capabilities of these pupils.
- The test was content valid as it applied to most levels and most parts of the National Curriculum Programme of Study.

- Over 45,000 pupils were awarded levels from the 2005 pilot. The level-awarding meeting took into account: the views of a teacher panel, the views of QCA ICT curriculum specialist and advice from the test developer, RM.
- Following the level awarding meeting, due diligence work was done to examine any cases of anomalous or potentially unfair (non-)awards.
- Substantial work is underway to review 2005 level awarding and to make sure that future awarding procedures are robust and defensible.

Areas for further work

- There is evidence from several sources that stakeholders did not feel that the test was face valid for level six.
- The initial measure of how consistently the test classified pupils into the same National Curriculum level on repeat administration showed that the test was approximately equally likely to classify a pupil into a different level, as to award that pupil the same level on repeat administration.
- Informed observers would have more confidence that the classification consistency of the test was reliable, if a large sample was achieved for a bespoke test-retest reliability study.
- As well as collecting a larger sample of empirical data, the project's attitude to and expectations of classification consistency should be clarified in future validation work.
- The reliability of test outcomes at level six was much lower than at other levels. This backs up other findings that cast doubt upon the validity of the test at level six.
- The methodology for comparing the test outcomes for pupils who spoke English as an Additional Language (EAL) with those who spoke English as their first language was less than ideal in that it combined all EAL pupils into a single, and by implication homogenous, group. It will be important to carry out some more broadly-based research to find out the fairness of the test for different sub-groups of pupils who do not speak English as their first language.
- The analysis of outcomes which took into account the largest variety of types of Special Educational Needs (SEN) (school action and pupils with statements), and a large sample of pupils, suggested that this test was not fair for pupils with SEN.
- Pupils using an 800*600-pixels screen resolution monitor appeared to be disadvantaged, as opposed to those who used a 1024*768 monitor. This issue should be further investigated.

- Pupils who sat both test sessions with a gap of six days or fewer scored more highly than those who had a longer time gap between sessions. The implications of this finding are not immediately clear. This analysis should be replicated in subsequent years and findings interpreted in the light of new results.
- There were some aspects of the test that were not sufficiently content valid: this applied particularly to level six and to aspects of the curriculum that covered the 'Communication' part of ICT.
- Some of the methods for evaluating content validity need to be improved for future years (this applies especially to opportunity counting).
- A study comparing teacher assessment and test outcomes did not provide concurrent evidence of validity for the Key Stage 3 ICT tests.
- It will be important, in describing the validity of the 2006 pilot, to be able to refer to a well-designed and executed concurrent validity study.
- The levels awarded by the ICT test were low, when compared with other NC tests, and compared with teacher assessment in ICT.
- There was a high percentage of pupils who were awarded no level from the test.
- At the present time, neither the project nor the formative evaluation can provide a definitive reason for the differences between 2005 test and TA outcomes.

Objective two: Infrastructure software and support processes scalability

Positive outcomes

- A high standard of support was provided across a large majority of service calls during the 2005 pilot.
- The project has identified the main peaks in service call volumes during the reporting period than ran from August 2004 to July 2005. This understanding of these potential causes of high call volumes, and the ongoing active drive to get schools to progress through accreditation, software installation and similar activities should help to increase the scalability of the solution for future years.
- A set of main lessons learned from the 2005 pilot has been identified. Focus on these and other lessons should improve the scalability of the support services.
- There is evidence that once a school has participated in a pilot, it seems to require less support for subsequent years. If 2005 schools require less support in subsequent years, it will be easier to provide scalable service.

Areas for further work

- Some Service Level Agreements (SLAs) for dealing with service calls were not met. Whilst the potential impact of such breaches on a larger cohort of participating schools is not clear, it is clearly desirable that SLAs are met.
- If the requirement for support were to increase in direct proportion with the number of schools to be involved in future pilots, then it might be difficult to provide a scalable service.

Objective three: Accurate formative and summative reports

Positive outcomes

- A set of detailed 'lessons learned' has been agreed between RM and QCA. It is intended that these lessons will lead to an improvement in the reports for 2006.
- QCA will focus on the formative use of e-assessment in a specific research activity in 2006, in order to improve the quality of formative reports.

Areas for further work

- Few teachers were believed to have seen the formative reports in the 2005 pilot.
- The formative reports received a poor approval rating from those few teachers that had seen them.
- QCA instructed RM not to release summative reports to schools along with test results. This was due to concerns about their fitness for purpose, and lack of usefulness to schools.

Objective four: Test security

Positive outcomes

- Certain features of the test (for example, cloning and randomisation) make it likely that copying from each others' screens will not be possible.
- Institutional security breaches that occurred during the 2005 pilot were not sufficiently serious to overturn the finding that this aspect of objective four had been passed.

Areas for further work

- The copying trial deployed as part of the 2005 validation did not have a sufficiently robust or valid methodology to demonstrate that the pupils could not copy from each others' screens.

- Four institutional security breaches were known to have occurred during the 2005 pilot.
- If the four security breaches had occurred in respect of statutory test material they would have had to have been taken more seriously.

Objective five: School experience

Positive outcomes

- There were some schools that, although they did not proceed fully with the 2005 pilot, still had a positive view of the software and overall experience.
- A new initiative of the Participation Task Force has been set up with the underlying purpose of ensuring that all schools are prepared for the 2008 test.

Areas for further work

- Although over 2,400 schools expressed initial interest in taking part in the 2005 pilot, only 402 sent back valid data from the first summative test window.
- Substantial numbers of schools ceased involvement with the pilot at several stages of the 2004 – 2005 cycle.
- The reasons for schools ceasing to be involved were many and varied, but some initial analysis has been done to investigate these reasons, and this could be followed up in subsequent years.
- Only a small number of schools ran full test sessions in a second window that was made available. No investigation has been run into the low uptake in the second window, nor has the usefulness of a second test window been evaluated.

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 the final evaluation report in that it considers all evidence that is available on the 2005 pilot.
3. This report describes the definitive findings of the formative evaluation of the 2005 pilot of the Key Stage 3 ICT tests. As such, it evaluates the pilot against:
   - the objectives of the 2005 pilot.
   - the objectives of the project overall, given that 2005 represents an intermediate staging post on the road to full statutory roll-out in 2008.
4. This report supplements the interim evaluation report, dated 12th July, which was delivered to the DfES. In order to supplement the interim report, the final report considers aspects of objectives that were not covered in the earlier document. It does not revisit aspects of objectives that were addressed in the 12th July report unless substantial new information has become available in intervening months.
5. Unless explicitly updated in the final report, findings of the interim report stand in their own right, and should not be seen as provisional or having less weight.
6. The second bullet point in paragraph 3 – the evaluation of the pilot overall – is an addition as compared to the interim report. It reflects the final, rather than interim nature of this document.

2.2 Structure of this report

7. The next section of this document (p. 8) lists all the sources of evidence that have been used to inform this report.
8. Following the evidence section, there are five sections that provide evaluation with respect to each separate 2005-pilot objective.
9. The final substantial section of this report (p. 53) evaluates the 2005 pilot against the overall aims and purpose of the Key Stage 3 ICT test development project.
10. This report does not give a detailed description of the Key Stage 3 ICT tests, or the associated test development project or the Key Stage 3 ICT assessment programme. Readers wishing such a background description are referred to Annex A of the interim evaluation report (interim report – p. 34).

3 Evidence

11. The interim evaluation report was based on many, but relatively provisional, sources of evidence (see interim report, page 6).
12. The current report is based on fewer formally-distinct sources of evidence, but these documents by and large represent the formally-stated, final position of the test developer – Research Machines PLC (RM). Further, such formal reports are generally based on a broad range of sources of evidence themselves.
13. Documents that have informed this report include:

- RM validity report
- RM level setting report
- Briefing notes sent to the QCA Chief Executive concerning the Key Stage 3 ICT tests
- DfES statistical first release showing the numbers of pupils awarded NC levels in 2005
- RM lessons learned report
- RM service delivery report
- Emails concerning non-release of summative reports
- RM 'test security issues 2005' document
- RM 'summative flightdeck' spreadsheet (tracking participation levels in the pilot windows)
- Key Stage 3 ICT tests Project Initiation Document
- Sundry email correspondence with relevant QCA and RM staff

4 Evaluation of 2005 objectives

4.1 Objective one

14. 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.

15. 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.

16. 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 Previous findings on objective one

17. Evaluation in the interim report addressed the following four areas:

- Test development procedures
- The developed test forms
- Qualitative findings from school visits
- Formal validation work

18. The 'formal validation work' section of the interim report was relatively limited, providing only the following information:

- A description of RM's validity specification and measurement model
- A table showing the level thresholds for the 2005 pilot

19. The interim report did not refer to formal RM reports based on summative data, nor to detailed analyses of distributions of awarded levels. This was because such reports had not been written at the time of the interim evaluation.

4.1.2 Bases for the validity findings in the final report

20. The validity section of this final report will be based around RM's formal validity report, the report of the level-awarding process, and the resulting distributions of levels.
21. In evaluating validity using RM's specification and report, this evaluation will consider the following aspects of validity:
    - Face validity
    - Reliability
    - Fairness for all pupils
    - Content validity
    - Concurrent validity
22. A definition of each aspect of validity from the preceding list will be given at the start of each relevant sub-section.
23. As well as analysing the list above of aspects of validity, the report will evaluate the pilot's level-awarding procedures, and the (distributions of) results that flowed from those procedures.
24. In addition to critiquing RM's validity findings from their reports, this evaluation will be informed by two specific principles set out in the highly regarded American publication *Standards for Educational and Psychological Testing* (AERA et al, 1999). That is:
    - Validity is most properly regarded as a unitary concept; that is, different 'types' of validity are better considered as different sources of evidence towards a single concept of validity.
    - The test developer and sponsor are under an active duty to provide strong evidence that the test is valid. If such evidence is absent, or questionable, then the best interpretation is that the test has not been demonstrated to be valid.
25. The preceding paragraph places substantial obligations upon those who wish to introduce this new test into the English education system. However, it must also be noted that this test is in pilot phase. 2005 was the first year that a convincing attempt at a level-awarding process was carried out, and it was also the first year that levels were returned to schools. The levels awarded were purposefully not held to be a baseline for future years (i.e. they were not considered to have 'set the standard' to which all future years' tests must be linked).
26. Thus, this evaluation must decide whether the 2005 pilot was or was not an effective pilot. It is not reasonable to judge the validity of the 2005 pilot as if it were a live National Curriculum test.

4.1.3 Findings

4.1.3.1 Aspects of validity

4.1.3.1.1 Face validity

27. Face validity is the extent to which a test (and its outcomes) is perceived to be accurate, appropriate and useful by non-technical users.
28. RM's validity report bases its face validity findings on evidence from a variety of sources. It concludes that:
    - There are grounds for believing that face validity has been established for the test as it relates to levels three to five.
    - There is evidence from several sources that stakeholders did not feel that the test was face valid for level six.

4.1.3.1.2 Reliability

29. Reliability is a crucial aspect of a test's validity. Whilst validity is the overarching concept '*is the test measuring what we intend it to measure?*', reliability is a narrower, but indispensable, aspect of validity. In effect, if a test is not reliable, it is not actually measuring anything at all.
30. From a more technical perspective, reliability can be defined as the consistency of measurements when a testing procedure is repeated on a population of individuals.
31. The definition in the previous paragraph emphasises reliability as robust measurement when a procedure is repeated. There are many methods for quantifying the degree to which a measurement procedure may be robustly replicated. No single method for estimating reliability is perfect.
32. The most natural and direct method for estimating reliability is to ask some pupils to sit the test twice. This method is known as 'test-retest reliability'.
33. There are several practical considerations that make test-retest reliability studies less attractive (e.g. finding enough pupils to do the test twice, the possibility that pupils' abilities or attitudes will change between the two test administrations). For this reason, many reliability studies (including those to derive reliability indices that are reported for current National Curriculum tests) employ a technique in which the internal data from a single administration of the test is divided many times to approximate a measure for the robustness of the testing procedure across two administrations. Such approaches to estimating reliability are known as internal consistency (of data) measures.
34. Despite their practicability, and widespread use, there are grounds for questioning the applicability of internal consistency measures of reliability in the

case of the Key Stage 3 ICT tests. This is because internal consistency measures are generally used with tests consisting of traditional items; it is not clear whether such reliability estimation techniques can be used with an opportunity-based test.

35. There are further, specific grounds to doubt that internal consistency techniques' appropriateness in the case of the Key Stage 3 ICT tests. This is because the levelled and structured data from opportunities would be likely to produce inflated figures for internal consistency measures.
36. Estimates of reliability should relate to a meaningful measure. In the present context, an explicit and direct estimate of the test's propensity to classify pupils into the same National Curriculum level on repeated administration would be more useful than a more abstract measurement of internal consistency of data generated in the test.
37. Estimates of reliability should be made across the range of ability reported by the test. In the current case, this means that if there are any differences in the reliability of classifications for different National Curriculum levels, this should be reported.
38. Reliability is reported on a scale from 0 to 1. There are no absolute, canonical values for what constitutes 'good' reliability. However, users of many high-quality tests consider reliability in excess of 0.8 to be acceptable.
39. RM's validity report findings on reliability can be summarised as follows:
40. RM conducted two test-retest reliability studies; firstly, comparing pupils' scores from the March pre-test with their summative test scores, secondly a group of pupils sat the summative test twice in a bespoke data collection.
41. 784 pupils were analysed in the pre-test-to-summative-test study (a reasonable size of sample). But, in the bespoke data collection there were fewer pupils (see paragraph 50 below). This smaller sample size reduced the usefulness of findings from this study.
42. The pre-test-to-summative-test study reported reliabilities in terms of 'opportunity counts'. It did so for opportunities overall, and for opportunities at particular levels.
43. The reliability coefficients for all opportunities were around 0.8 in all cases (between 0.769 and 0.803).
44. The coefficients for reliability of opportunity counts with respect to individual levels tended to be rather lower than the values for all opportunities. This was especially so in the case of level six opportunities, whose reliability went down as 0.344 on one occasion.

45. The bespoke data collection study returned reliability coefficients of 0.717 and 0.780 for the lower and upper tier of the test, respectively. Once again, values with respect to specific levels were lower, with level six reliability being by far the lowest.
46. RM also conducted an internal consistency analysis on the test data. This showed reliability coefficients of 0.788 for the lower tier and 0.799 for the higher.
47. As in the previous studies, reliability was lower in the case of individual levels, with level six reliability being very low – 0.122.
48. The quoted reliability indices demonstrate that this test is capable of providing measurement that approaches the lower bound of what is considered to be acceptable in high-quality testing.
49. There are reasons to believe that such reliability indices might increase once the test has bedded down more; for instance, there is evidence that pupils' ICT capabilities are not evenly distributed (they tend to be weak in modelling and data handling). Such uneven ability in an assessed subject will inevitably reduce the consistency of data produced by pupils in response to a test and hence the reliability.
50. Some initial analysis was carried out to demonstrate the classification consistency of the tests. This was based on the bespoke test-retest data. In this study, it was found that, from 64 pupils doing the study for the lower tier, 31 were awarded a different level on the second test sitting as opposed to the first. For the higher tier, 46 out of 86 pupils were awarded the same level in both test sittings.
51. Thus, an ‘indicative classification consistency’ for this test is about 50 per cent.
52. This would seem, on the face of it, a rather poor result. This may be so; however, it is not necessarily incompatible with classification consistency rates in other high-stakes assessments. For example, Wiliam (2000) has suggested that a reliability co-efficient of 0.8 on an internal consistency measure would lead to the misclassification of between 19 and 43 per cent of pupils, depending upon how many levels were being used. Royal-Dawson (2005) has studied the accuracy of level classifications by markers of English papers in a mark-remark study. The classification consistency was found to be around 50 per cent for several different types of markers (differing skills and experience).
53. It will be important that the project’s attitude to and expectations of classification consistency, as well as other facets of reliability, is clarified in future validation work.

54. Finally, in summary, there are some ways in which reliability analysis is not yet adequate.
55. The initial measure of how consistently the test classified pupils into the same National Curriculum level on repeat administration showed that the test was approximately equally likely to classify a pupil into a different level, as to award that pupil the same level on repeat administration.
56. The bespoke test-retest reliability study was hindered by being based on only a small sample of pupils. Informed observers of the test's reliability would have more confidence that the test was reliable, if a large sample was achieved for a test-retest reliability study.
57. The reliability of test outcomes at level six were much lower than at other levels. This backs up other findings that cast doubt upon the validity of the test at level six; from the interim report (that there was not enough level six material in the test – see paragraph 26 of the interim report), and from this report (that a variety of commentators did not feel the test to have face validity at level six – see paragraph 28 above).

4.1.3.1.3 Fairness for all pupils

58. It is important that the Key Stage 3 ICT test is equally fair for all pupils. An excellent definition of 'fairness for all pupils' was given by the independent panel of experts that investigated A Level standards in 2002:

> Fairness … addresses the question of whether students given the same quality of preparation and who have the same degree of motivation would be likely to perform similarly in the examinations in question. Fairness involves the extent to which the test administration and scoring practices are comparable across identifiable groups of students. … Our use of the term 'fairness' in this fashion is not intended to convey that the performances of particular subgroups should be more or less equal, although that use of the term is sometimes made. Differences in group performance may be due to differences in preparation, e.g. quality of teaching, access to support, motivation, as well as to any differences among the subgroups, such as English language proficiency. (International panel, 2002)

59. Fairness in the current context applies to pupils' demographic characteristics (gender, English as an Additional Language (EAL) status, free school meals (FSM) status, etc.), and to pupils with special educational needs (SEN). The fairness requirement also obliges the test to be fair to pupils taking the test in different circumstances (e.g. using computer monitors with different resolutions, at different times in the test window, and so on).
60. The requirement that the test be fair for the various groups of pupils described in the last paragraph does not, as the quote from the international panel shows,

mean that all the groups must score at the same level. Rather, it means that any differences must be proportionate, must represent their underlying abilities and be consistent with other information on groups of pupils' abilities.

61. RM's validity report findings on fairness for all pupils are summarised and evaluated in the following paragraphs.
62. In the level three-to-five tier of the test, boys' and girls' performances were very similar, although boys performed slightly better. Similarly, in the level four-to-six tier, boys performed slightly better than girls. However, the boys' abilities were more widely spread. These findings suggest that this test was fair for both genders.
63. The performance of pupils who were eligible for free school meals was compared to those who were not. Pupils who were eligible for FSM scored significantly less well on the test than those whose families paid for their lunches.
64. Poorer performance by pupils who are eligible for free school meals is (regrettably) a phenomenon that can be observed in many test data sets. For example, in 2004 public examinations and national tests, pupils not eligible for free school meals performed better than those who were eligible for free dinners in each Key Stage, at GCSE and equivalent and at Post-16.
65. No analysis has investigated whether the differences in scoring between non-FSM-eligible and FSM-eligible pupils in current NC tests and the ICT test are comparable (or whether there was a bigger or smaller gap in the new test). Further, it has not been investigated whether the differences in scoring were associated with any specific factors related to the new test (for example, an impact of differential access to ICT in the home on scoring in this Key Stage 3 ICT test).
66. Bearing in mind the two caveats in the previous paragraph, the lower scoring of pupils eligible for free-school meals seems consistent with patterns of scoring in many other tests. Therefore, it is reasonable to conclude that the lower scoring by pupils entitled to free school meals in the Key Stage 3 ICT tests reflected the genuinely lower capabilities of these pupils.
67. Analysis was conducted on the test data to compare the scoring of pupils who had English as their first language with 'others'. Unfortunately, treating all pupils who do not speak English as their first language as a single homogenous group is unconvincing: learners with EAL are a very diverse group of pupils, with very different first languages, competencies in English language generally and literacy skills particularly.

68. The issue of the test's fairness for EAL pupils is a live one. This is particular so in light of the finding of the interim evaluation report that many pupils struggled to understand task instructions (interim report – paragraph 32). It is therefore important that a credible methodology for investigating the performance of diverse sub-groups of EAL pupils is developed.
69. The analysis that was done on the pilot data showed that EAL pupils scored less well on the test overall than did those who spoke English as a first language. However, it was not known whether this discrepancy represented a genuine difference in ability between the two groups. Teachers' Initial Level Assessments (ILAs) confirmed the view that EAL pupils' ICT capabilities were, on average, lower than those of pupils who spoke English as their first language. ILAs had been shown elsewhere to not necessarily be a good indicator of pupils' ICT capability, however.
70. School staff were able to use administrative software to key in pupils' Special Educational Needs status, when entering them for the test. Initial counts of the number of pupils who had SEN indicated that there were relatively few pupils with SEN entered for the test (and/or that staff had not entered the SEN information into the administration software for all pupils).
71. Pupils with SEN scored less well on the test overall than those who did not have special needs. However, the small number of pupils indicated as having SEN in the KS 3 data meant that statistical comparisons of the two groups were not very secure.
72. Therefore, some further analysis was done. DfES data showing pupils' SEN status were obtained. These data included information telling whether pupils had statements of special needs, but also whether pupils were subject to the three gradations of school-based special needs provision.
73. Then, the mean values of ILAs attributed to pupils (SEN and not) were calculated. Also, the mean levels that all pupils (SEN and not) were awarded were calculated. The hypothesis underlying this analysis was that the levels that SEN pupils were awarded should have a similar relationship to ILAs to that of pupils without SEN (i.e. it might be the case that all pupils' level awards were deflated as compared to ILA, but pupils with SEN must not have especially deflated awarded levels).
74. The results of this analysis are shown in Table 2:

	DfES SEN status	Mean	N	Std. Deviation
ILA	N: no special provision	4.80	39,128	0.72
	A: school action	4.36	3,778	0.76
	P: school action plus	4.23	1,344	0.81
	S: school action plus and statutory assessment	4.07	706	0.85
	Total		44,956	
Level	N: no special provision	3.77	39,128	1.72
	A: school action	2.53	3778	1.88
	P: school action plus	2.40	1344	1.91
	S: school action plus and statutory assessment	2.17	706	1.92
	Total		44,956[1]	

Table 2: ILAs and awarded levels for pupils with special needs

75. Table 2 shows that the awarded levels for pupils without SEN were lower than their ILAs. However, the awarded levels for pupils who were subject to the three types of school action, and pupils with statements, were much lower than their ILAs, and disproportionately lower when compared to pupils without SEN.
76. This finding suggests that this test was not fair for pupils with Special Educational Needs.
77. The original specification for the Key Stage 3 tests stated that tests could only be taken on monitors with a 1024*768-pixel[2] or larger resolution. 1024*768 resolution has the potential to display more information on screen than 800*600 resolution – either displaying more small text (and graphics) than a 800*600 resolution or displaying the same amount of information but at a higher resolution (quality).
78. This is important, since the screen for the Key Stage 3 ICT tests contains quite a lot of information, and has even been said to be 'busy' or 'cluttered' (see interim evaluation report – paragraph 40).
79. However, in order to allow as many schools as possible to participate in the test pilots, the minimum specification for screen resolution was reduced to 800*600 in a temporary change request.
80. Researchers compared the scoring of pupils who used 800*600 resolution settings with those who used 1024*768 resolution settings. Substantial groups of

[1] It was not possible to find matching SEN data for all the pupils who took part in the pilot.

[2] Pixel can be defined as follows: a combination of the words 'picture' and 'element'. A pixel is the smallest discernible sample of video information, the 'little squares' that make up an overall picture.

pupils used the respective monitor resolutions; and comparisons of scores for the upper and lower tiers of the test showed a sustained pattern of lower scoring for pupils using the lower monitor resolution (800*600).

81. Thus, the use of a smaller-resolution monitor appears to disadvantage pupils. The directness (and size) of the effect of monitor resolution, and its link to other factors (such as the school's overall quality of ICT kit) are not clear, however. This issue should be investigated further.
82. Schools were able to timetable the test within a four-week window. Analysts sought to establish whether there was any effect for pupils doing sessions in different weeks and for pupils having differing gaps between their test sessions.
83. The most striking findings with respect to timetabling test sessions within the four-week window relate to the gap between sessions. Here, it was found that pupils sitting session two six or fewer days after session one tended to score more highly than those who sat the two sessions with a longer time gap.
84. It would be useful to replicate such analyses on future years' test data. Also, there could be several potential implications of this finding; for example, it may be that sitting both sessions close together provides an advantage to pupils, whilst having the option to timetable sessions throughout the four-week window could be an important convenience for teachers. If this was the case, it might be necessary to resolve an important clash of priorities.

4.1.3.1.4 Content validity

85. Content validity can be defined as: 'whether a test adequately targets and represents the whole performance domain'. In the current case, the whole performance domain is the National Curriculum for ICT, and it is agreed within the project that the QCA Rules Base represents the most legitimate device for operationalising the curriculum in an e-assessment, and thus that coverage of the Rules Base is an important factor to consider when evaluating content validity.
86. The RM validity report analyses the content aspect of validity by:
    - counting the numbers of elaborations in the test at different levels.
    - asking teachers at a review group how much of the curriculum they perceived the test to cover.
    - counting the numbers of opportunities in the test that mapped to different aspects of the curriculum Programme of Study (PoS), and to sub-divisions thereof (i.e. ICT capabilities).
87. Findings from these analyses included the following points.
88. More than 70 per cent of available Rules Base elaborations were included in the tests at all National Curriculum levels (three to six).

89. Analysis of the Teacher Review Group's (TRG's) perceptions of the areas of the programme of study that were covered in the test shows that the teachers thought that 81 per cent of the PoS was covered in the test.
90. The table below has been constructed to show those aspects of the Programme of Study that the TRG did **not** perceive to be covered, and possible reasons for this perception:

Knowledge, skills and understanding	**Aspect of PoS (*Pupils should be taught:*)**	**Possible reasons for perceptions of non-inclusion**
2. Developing ideas and making things happen	b. to recognise where groups of instructions need repeating and to automate frequently used processes by constructing efficient procedures that are fit for purpose.	2005 test deliberately did not address topic of 'control'.
4. Reviewing, modifying and evaluating work as it progresses	a. reflect critically on their own and others' uses of ICT to help them develop and improve their ideas and the quality of their work.	It may be difficult to assess 'reflecting critically on ideas' in a timed test.
	b. share their views and experiences of ICT, considering the range of its uses and talking about its significance to individuals, communities and society.	It may be difficult to authentically assess 'sharing views and experiences' in a test in which traditional examination conditions apply.
	c. discuss how they might use ICT in future work and how they would judge its effectiveness, using relevant technical terms.	It may be difficult to realistically discuss the use of ICT in future work in a test in which traditional examination conditions apply.
5. Breadth of study	***During the key stage, pupils should be taught the knowledge, skills and understanding through:***	
	b. working with others to explore a variety of information sources and ICT tools in a variety of contexts.	It may be difficult to authentically assess 'working with others' in a test in which traditional examination conditions apply.
	d. comparing their use of ICT with its use in the wider world.	This aspect may require a longer written response.

Table 3: Aspects of the PoS perceived by teachers to not be covered in the test

91. In discussing the implications of this table, it is important to spell out two important limitations to the generalisability and reliability of findings:
    - This table is based upon the opinions and perceptions of one group of teachers. Although this group was diverse, and selected as a national group, these views might not be representative of the opinions of all teachers in the pilot.
    - Whilst the minutes of the TRG meeting that led to the discussions about curriculum coverage have been studied, the right-hand column in the table

has in fact been added by the writer of this evaluation report. As such, it represents informed but limited judgement, rather than hard data.

92. Whilst accepting that the previous paragraph puts limitations on interpretations of Table 3, it is possible to make some relevant observations:

    - Several of the non-covered aspects of the PoS relate to the 'Communication' part of ICT. A teacher comment from the Review Group minutes puts this eloquently:

      [It is] hard to see how 'Reviewing, modifying and evaluating' can be covered … many of the statements include the terms 'share', 'discuss', and 'reflect'. Similarly with 'Breadth of Study', 'Working with others'. Generally it's about the 'C' in ICT and how this is assessed.

    - The table contains aspects of the PoS, which is not levelled. However, some of the aspects that the teachers perceived not to be covered had close parallels in the higher levels of the NC level descriptions (especially in level seven). This is particularly true of aspect 4a (*reflect critically* …). In this sense, it may be that some of the non-covered aspects related to levels seven and eight; this may have explained why they were not covered in this level three-to-six test.
    - It is reasonable that teachers should perceive the 2005 test to not address 'control', since it was not designed to do so.
    - It is rather less reasonable that teachers perceived the test to not facilitate the assessment of pupils' ability to compare ICT use with that in the wider world. The reasons for this perceived omission should be further investigated.

93. Counts of opportunities and their mappings to the nine ICT capabilities suffered from a methodological weakness, in that opportunities were often mapped to multiple capabilities and there was a possibility that this could lead to undesirable 'double counting'.

94. Bearing in mind this limitation in the data analysis, the following findings can be reported:

    - Opportunity counts showed the majority of ICT capabilities to have been covered in both tiers of the test.
    - There were few opportunities for pupils to present information or review work in the lower tier of the test.
    - In the upper tier, level six had fewer opportunities than other levels.
    - Within the generally weak coverage of level six, 'reviewing modifying and evaluating work as it progresses' was especially sparsely covered.

95. In considering content validity, it is also important to reiterate two findings from the interim report (paragraphs 26 – 30):

    - Employing the opportunity-counting method that was used before the release of the 2005 tests, there appeared to be far fewer level six opportunities than there were opportunities at other levels.
    - The vast majority of level-six opportunities in the test were available in the last task, which was based on data handling.

96. Thus, the findings on content validity are quite diverse. However, they may be summarised as follows:

    - There is good evidence, from a variety of sources, that the test was content valid as it applied to most levels and most parts of the Programme of Study.
    - There were some senses in which the test was not content valid: this applied particularly to level six and to aspects of the curriculum that covered the 'Communication' part of ICT.
    - Some of the methods for evaluating content validity could be improved for future years (this applies especially to opportunity counting).

4.1.3.1.5 Concurrent validity

97. Concurrent validity can be defined as: 'the extent to which the outcomes of an assessment are consistent with other independently obtained measures of the construct or performance of interest'.

98. RM's validity report contained three potential independent measurement procedures that could have formed the basis for concurrent validity studies:

    - Some skilled observers from the Centre for Formative Assessment studies (CFAS), University of Manchester watched pupils' tests and then assigned the pupil to a level.
    - Some skilled moderators from CFAS reviewed opportunity description reports[3], and assigned pupils to a level.
    - Teacher assessments (TAs) were available for a sub-set of the pupils who took the test (approximately ten per cent of all pupils in the pilot).

99. Unfortunately, neither of the first two methods provided good data on which to base a concurrency study; this was either because of small sample sizes or because there were some lessons to learn in terms of methodology (e.g. judges being unsure about specific details of how to assign pupils to levels, or how to record levels).

100. Therefore, the concurrency study in the 2005 pilot was based on teacher assessment alone. Unfortunately, this study did not provide evidence of concurrent validity, as there was not much agreement between the level awarded in the test and the level awarded by TA. (For further description of intrinsic discrepancies between test results and TA see paragraph 118 below.)

101. It will be important, in describing the validity of the 2006 pilot, to be able to refer to a well-designed and executed concurrent validity study.

[3] During the KS3 ICT test, pupils display evidence of ICT capability. This is captured in terms of opportunities. Each opportunity is described in language that a person familiar with the ICT curriculum would understand. Opportunity descriptions can be concatenated to form a report – effectively listing all the evidence of ICT capability that each pupil demonstrated in the test.

102. Proposed work in 2006 includes establishing the most appropriate and practical measurement procedures with which to conduct a concurrent study, and then the conducting of this study.

4.1.3.2 Level setting

4.1.3.2.1 Level setting procedures

103. Level awarding in the 2005 pilot was conducted using the sufficient evidence model. This measurement and awarding model has been specifically developed for the Key Stage 3 ICT tests (although it belongs to an established family of level-awarding models).

104. Features of the sufficient evidence model as operationalised in 2005 awarding included:

- In order to be awarded a National Curriculum level, pupils had to gain a specified number of opportunities that were targeted at that level. Opportunities were 'targeted at a level' in that they related to elaborations, and through the QCA Rules Base, they could be ultimately linked to National Curriculum level descriptions.
- In order to be awarded a level that was not the bottom of a tier (i.e. levels four and five in the three-to-five tier and level five in the four-to-six tier), pupils had to demonstrate a certain number of opportunities at the level to be awarded, and also had to demonstrate that they would be awarded the level(s) below as well.
- Level six awarding was performed on a different basis to other levels in 2005. To be awarded level six, pupils had to demonstrate that they were a 'sound level five' – that is, that they had gained a number of opportunities well above the level five cut score – and they also had to have gained a small number of level six opportunities (in fact, the number was one).

105. A panel of teachers and QCA's ICT curriculum specialist were separately shown opportunity description reports (see footnote 3, above, at page 21). They were then taken through a procedure which output recommended numbers of opportunities that would be necessary for pupils to gain, in order to be awarded the different levels (cut scores).

106. RM analysts provided a form of high and low score for opportunities at each level. They did this, respectively, by:

- counting the number of opportunities that would have been taken by a pupil who had taken all the available opportunities at a given level in the test.
- counting the number of opportunities that would have been taken by a pupil who had shown ICT evidence of each particular type once, but only once.

107. Two level-awarding meetings were convened. The first meeting considered descriptive information on the 2005 administration, and was chaired by RM. The second meeting made decisions about cut scores, and was chaired by QCA. The meetings were attended by staff from QCA, RM and CFAS.

108. The meetings were presented with general analysis based on the pilot (raw counts of opportunities, performance of different demographic groups, etc.), and with the three alternative sets of cut scores described in paragraphs 104 – 106 above.

109. The level awarding meeting adjudicated between the three potential sets of cut scores, and decided on a definitive set of cut scores that was used to set levels for the 45,000 pupils. The awarding meeting was given initial feedback as to the likely distribution of levels, given the cut scores that had been decided upon.

110. Following the meeting, RM and CFAS researchers carried out due diligence checks to make sure that the level-setting procedure had been based upon accurate data, and that there was not an unacceptable number of anomalous and potentially unfair results (e.g. large numbers of pupils who had demonstrated sufficient evidence at a level, but who had not demonstrated evidence at the level below).

4.1.3.2.2 Distributions of awarded levels

111. The agreed cut scores were reported in the interim evaluation report (Table 3, page 14).

112. The numbers of pupils (and accompanying percentages) that were awarded each level are shown in the following table:

National Curriculum level awarded	Number of pupils	Percentage	Cumulative percentage
'n'	7,715	16.9%	16.9%
3	6,066	13.3%	30.3%
4	15,332	33.7%	63.9%
5	13,731	30.2%	94.1%
6	2,696	5.9%	100%
Total	**45,540**	**100%**	

Table 4: Distribution of pupils awarded NC levels in the 2005 pilot

113. Following awarding, due diligence was performed to make sure that there were not significant numbers of pupils who had received an unfair level award as a result of an unforeseen consequence of the sufficient evidence model. In particular, analysts concentrated on the possible situation in which pupils had

scored enough to have been awarded a level that was not the bottom of the tier, but who had not scored enough opportunities from the level below to be awarded the level (e.g. a pupil in the lower tier who would have had enough level four opportunities to be awarded level four, except that she did not have enough level three opportunities to be awarded that level).

114. Findings of the due diligence investigations for anomalous performance included:

- In the lower tier there were no pupils who would have achieved a level if the condition requiring achievement of the lower level had been removed.
- In the lower tier there was a small number of pupils who had showed some evidence of level four, but who were awarded no level (two per cent of all pupils who were awarded a level 'n').
- In the upper tier there was a small group of pupils who would have achieved a level five if the condition to achieve a level four had not been in place (0.2 per cent of the pupils who were awarded a level 'n').
- There was also a group of pupils who showed significant, but not sufficient, evidence of level five, but who were awarded no level (three per cent of the pupils who were awarded a level 'n').

115. Sophisticated analytical techniques (using box plots) to identify potentially anomalous (non-)awards were developed in 2005, and it is suggested that these are given wider operational use in 2006.

4.1.3.2.3 Discussion of distributions of awarded levels

116. The levels achieved by pupils in the 2005 pilot were low. This is true when compared to pupils' achieved levels in other National Curriculum tests and in comparison with levels achieved from ICT teacher assessment.

117. A comparison of levels achieved in ICT via the test and via TA is given in the figure below:

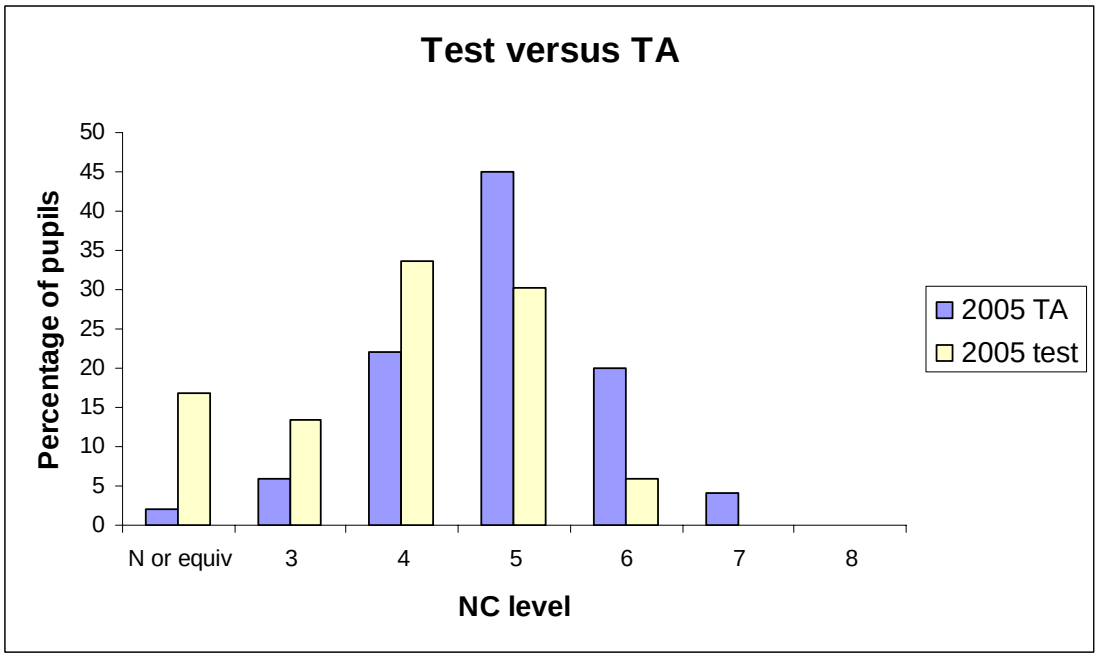

Figure 1: Distributions of test and Teacher Assessment results

118. Comparison of the distributions of ICT test and TA results from the graph above shows two features:

- Levels awarded by TA are approximately one level higher than the levels awarded by the test – i.e. the distribution of test results has shifted one level 'to the left' when compared with the TA distribution.
- There is a much larger proportion of pupils who have been awarded no level ('level n') from the test as compared to TA.

119. There are several potential reasons for the differences between test and TA results.

120. The issue of the TA results being approximately one level higher is discussed first of all.

121. Previous years' Ofsted reports have given cause to believe that teacher assessment in ICT may have been systematically either one or two levels more generous than should have been merited by pupils' ICT capabilities.

122. However, the current evaluation has sought the opinions of national authorities on ICT (the director of the ICT strand of the secondary strategy, QCA ICT curriculum consultants, and the then HMI, Specialist Adviser for ICT, Ofsted). The consensus of those experts is that TA is more accurate than it used to be, and unlikely to be as much as one level too high.

123. Other possible causes for the discrepancy in results between the test and TA include:

- The test and TA measure slightly different things and therefore it might be legitimate for there to be a difference in the levels awarded by these two distinct measurement procedures.
- The joint facts that the test was novel and in pilot phase may have caused pupils to perform less well than they were actually capable of (e.g. if pupils knew that this was only a pilot and, so, did not try their hardest, or if they were not yet sufficiently acquainted with the test toolbox).
- There may be some illegitimate source of difficulty in the test which caused pupils to perform less well than they were capable of.

124. Similarly, there may be several causes for the high proportion of pupils in the test who were awarded a level 'n'. Potential causes include:

- The fact that many pupils seem to have been inappropriately entered for the higher tier – hence they ended up receiving a level 'n', when they might more properly have been awarded level three.
- Pupils' unfamiliarity with, or possible lack of commitment to, the test causing them to score less well than their actual capability would merit.
- Some technical flaw in the test or measurement design that tends to cause an unreasonably large proportion of tests to receive a level 'n' result.

125. Anecdotal reports of the reaction of the teachers rating tests using opportunity reports in the initial phase of the awarding process were that they were comfortable with results from the level three-to-five tier of the test, but not from the higher tier. (In the higher tier, their assessment was harsher than the cut scores eventually agreed upon.)

126. At the present time, neither the project nor the formative evaluation can provide a definitive reason for the differences between test and TA outcomes.

127. Current and future work includes several substantial activities that should make sure that awarding procedures are robust and defensible. Such strands include: an independent review of the 2005 pilot by an expert panel, and some sustained and detailed research into sources of difficulty in the test.

128. The focus of the work described above will be to assure the validity of awarding procedures and of the test in general; it may well not be fruitful (nor, indeed, possible) to directly answer the question 'were the 2005 TA or test results the correct measurement of pupils' ICT capabilities?'

129. Rather, validity work will (to simplify considerably) seek to establish whether the levels awarded by the test suggest that the test is assessing ICT capability, and that it is doing so robustly.

4.1.3.3 Delivery of results data to the DfES

130. Test data were sent to the DfES on 9th September 2005, following an apparently successful dummy run in the summer of the same year.

131. The DfES has confirmed that the data was received on 9th September 2005, and that the transfer was made using a suitably secure method. An initial eyeballing of the data set showed it to be in an appropriate format.

4.1.4 Evaluation of objective one

132. The interim evaluation reported findings in the categories:

- Test development procedures
- Developed test forms
- Qualitative findings from school visits

133. For the purposes of the current report, validity has been taken to have several aspects (see paragraph 21 above). Also, evaluation of validity has considered level awarding, and the delivery of results to the DfES.

134. Whilst looking at those different aspects of the construct, the position of the evaluation is that validity is a unitary construct, and that a single statement should be made as to whether the test is valid or not (see paragraph 24).

135. A further position of this evaluation is that the 2005 pilot will be judged as a pilot – not as if it were an administration of a current statutory test (see paragraphs 3 and 256).

136. Many positive findings have been reported in the evaluation sections of the interim and final reports. However, the following table looks at each aspect of validity, states whether the aspect has been achieved, and summarises areas that still remain to be proven.

Interim report or final?	Aspect of validity	Validity established?	Outstanding areas where validity must be proven
Interim	Test development procedures	Partly	• Some lessons still to learn to improve quality.
	Developed test forms	Partly	• Different numbers of opportunities in two tiers at level five • Too few opportunities at level six
	Qualitative findings from school visits	Mostly	• Task instructions could be difficult to comprehend • Some pupils felt screens to be cluttered
Final	Face validity	Mostly	• Users did not perceive to be valid for level six
	Reliability	Not clearly established	• Need for single indicator of the reliability of level classifications. • Reliability lower at level six.
	Fairness for all pupils	Established with some exceptions	• Pupils with EAL • Pupils with SEN • Effects of monitor resolution and gap between test sessions – clarification needed
	Content validity	Mostly	• Level six content • Communication aspects of ICT
	Concurrent validity	Not at all proven	• Convincing concurrent validity study necessary
	Evidence of valid level setting	Procedures awaiting validation	• Basis for awarding level six different to other levels • Low distribution of awards • Large proportion of level 'n' awards
	Test data delivered successfully to the DfES	Yes	

Table 5: Aspects of validity: whether validity established and outstanding areas

137. The table shows that there remain a significant number of major tasks still to be undertaken to demonstrate validity. These tasks can be summarised as follows:

- Improve some test development procedures
- Demonstrate that tests are composed of suitable amounts of material
- Demonstrate that the (lack of) readability does not impede the assessment of pupils' ICT capabilities
- Increase level six content in the test
- Undertake major studies to derive formal indices for reliability (classification consistency) and (concurrent) validity

- Establish the test's fairness for substantial demographic groups (pupils who speak English as an Additional Language, and pupils with Special Educational Needs)
- Confirm that level-awarding methods are effective.

138. The amount of work implied by the list should not be underestimated. Also, the 'areas for further work' listed in the executive summary of this report should form the basis for further validation. Nevertheless, it does seem that the major outstanding tasks for the demonstration of validity are reasonable for the project to achieve before the test is administered as a statutory National Curriculum test.

139. The 2005 pilot has therefore achieved objective one. If the project continues a sustained programme of work to improve quality, it is likely that the test will be sufficiently valid to be released as a statutory test in 2008.

4.2 Objective two

140. 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.

141. 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.

142. 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.

143. This CSF was written before the amendment to objective two described in paragraph 141. Therefore, it only refers to the first half of the amended objective.

144. 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 Previous findings on objective two

145. The interim evaluation found that the infrastructure software reliability and scalability aspect of objective two had been passed. It did not comment on the scalability of technical and customer services support.

146. Hence the rest of this section of this final evaluation report will concentrate on support services.

4.2.2 Findings

4.2.2.1 Technical and customer services support

4.2.2.1.1 Types of support provided

147. Support to schools participating in the 2005 pilot of the Key Stage 3 ICT tests was split into two aspects: customer services and technical support.

148. Support to schools was multi-faceted. Communication channels available to provide support included: a phone line, an email address and a web site.

149. Types of support included (amongst many other things):

- online and paper copies of various manuals (e.g. installation manuals)
- online demos of the software and tutorials in how to use it
- posters
- knowledge-based articles relating to known features of the test software
- Frequently-asked questions (FAQs)

150. The suite of support materials was designed to serve a range of different purposes (e.g. quick-start guides and detailed manuals) and to have features to aid use (e.g. indexes and internal search facilities).

151. The intention of the support model was that it would be mainly conducted remotely. There were conferences for Local Education Authority staff in October 2004, to convey key messages about the test. Also, on a few occasions RM support team staff visited schools to clear up problems. But the intention was that most support would be provided via the remote channels described in the foregoing paragraphs.

4.2.2.1.2 Success against contractual measures for service calls

152. Service calls are all communications from schools that contain a substantial issue that was logged (e.g. more than just a phone call to say 'thank you'). Service calls can be generated from several communication modes, for example: from a telephone call, email or via a member of a school's staff filling in a form in Support Online.

153. A clause in the contract between QCA and RM imposed a Service Level Agreement (SLA) with respect to the speed of resolution of service calls. The SLA divided calls into three degrees of severity: critical, material and cosmetic. It also required the service provider to implement a temporary solution to a problem within a specified timeframe, and to effect a full solution within a longer time period. Finally, the SLA required that certain percentages of all calls of each severity be resolved within specific timeframes (e.g. 80 per cent of all critical calls should be provided with a temporary solution within four hours, which should then be fully corrected within two weeks). The exact parameters of the SLA are shown in the following table:

Fault	Temporary solution	Correction within
Critical	80% - 4 hours 95% - 1 day 100% - 1 week	2 weeks
Material	80% - 1 day 95% - 2 days 100% - 1 week	1 month
Cosmetic	2 weeks	1 month

Table 6: Parameters for service calls SLA

154. The success or otherwise of the service provision with respect to its SLA targets was evaluated. The success rates for service provision with respect to the various SLA criteria are described in the tables below. Table 7 shows the total number of service calls and numbers of service calls that did not meet various SLA criteria. The short titles in the table columns can be expanded as follows:

- Number of calls (2005): the total number of technical service calls logged by Technical Support, which were monitored according to the contracted SLA
- >80% Temp Solution: the number of calls for which no temporary solution was provided within the elapsed time that should have been achieved for 80 per cent of calls (i.e. four hours for critical calls, one day for material calls and two weeks for cosmetic calls)
- >95% Temp Solution: the number of calls for which no temporary solution was provided within the elapsed time that should have been achieved for 95 per cent of calls
- >100% Temp Solution: the number of calls for which no temporary solution was provided within the elapsed time that should have been achieved for 100 per cent of calls
- >Correction: the number of calls that did not receive a full solution with the elapsed time period that was appropriate to the degree of severity of the call (i.e. two weeks for critical, and one month for both material and cosmetic calls).

Call Severity	Number of calls (2005)	>80% Temp Solution	>95% Temp Solution	>100% Temp Solution	>Correction
Critical	113	21	14	11	5
Material	1684	111	75	46	9
Cosmetic	207	1	1	1	1
Total Calls	**2004**	**133**	**90**	**58**	**15**

Table 7: Numbers of service calls that did not meet SLA targets

155. A follow-up table has been developed to further interpret Table 7. Table 8 takes the raw number of calls from its predecessor table and provides them in the form of percentages. So, for example, there were 21 calls that did not meet the first threshold for critical calls, from a total of 113 calls. This amounts to 18.58

per cent of calls at that threshold and severity. Further, in Table 8 shading shows where a threshold has not been achieved. For example 21 critical calls were not solved within one day. This amounts to 12.39 per cent of all such calls (113). This is more than the relevant threshold (five per cent) and the cell is therefore shaded in the table.

Call Severity	Temporary solution 'Failure thresholds'			>Correction
	20%	5%	0%	
Critical	18.58%	12.39%	9.73%	4.42%
Material	6.59%	4.45%	2.73%	0.53%
Cosmetic	0.48%	0.48%	0.48%	0.48%
Overall percentages	**6.64%**	**4.49%**	**2.89%**	**0.75%**

Table 8: Percentages of service calls that did not meet SLA targets

156. Table 8 permits the following observations.

157. Temporary solutions were effected within the timeframe for all severities of calls for the 80 per cent threshold (i.e. less than four hours for critical calls, less than one day for material and less than two weeks for cosmetic calls).

158. This is important, since this was the biggest group of calls, and the requirement was the most stringent (shortest timeframe).

159. The 95 per cent threshold was satisfied in the case of material and cosmetic calls, but not for critical calls.

160. There were some calls of each severity that were not given a temporary solution within one week (critical and material calls) or within two weeks (cosmetic calls).

161. Similarly, there were 15 of the 2,004 calls received during the pilot for which no long-term correction was applied. All of these calls were classified by RM as Issue Centric Calls requiring significant investigation to diagnose and resolve the underlying issues. All of these issues are being addressed for the 2006 pilot.

162. Overall, these tables show that a good level of service was provided, but that SLAs were breached in several areas. Whilst the potential impact of such breaches on a larger cohort of participating schools is not clear, it is clearly desirable that SLAs are met.

4.2.2.1.3 Factors that affected service call volumes

163. The volume of service calls received was not uniform across the project. There were several peaks in the numbers of service calls received. The five most prominent peaks in activity are displayed in the following figure.

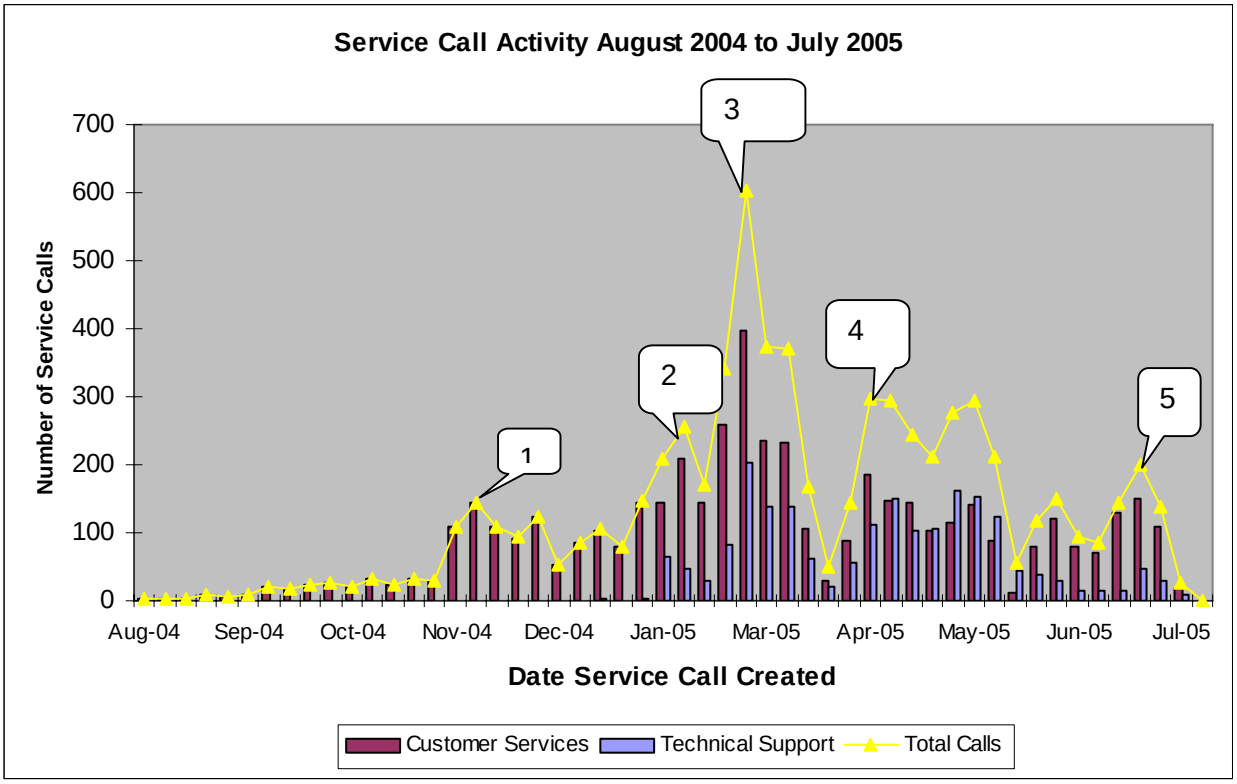

Figure 2: Peaks in the number of service calls in 2004 – 2005

164. The five peaks can be explained as follows:

1. *Accreditation process*: schools needed to be accredited to take part in the pilot. Substantial aspects of accreditation included running the network audit tool to make sure that schools' computers were technically adequate to take part in the project, and getting schools' senior managements to sign a commitment form. It was necessary for service teams to prompt schools to progress through these activities. This increased the volume of service activity.
2. *Release of 2005 software*: Technical support had to assist with installation queries. Thus, the number of technical support calls increased at this time, although there were still fewer technical than customer services calls at this stage.
3. *Installation deadline:* The volume of both customer services and technical support calls peaked at the end of February and beginning of March. This was due to the 28th February deadline for installing infrastructure software. 321 service calls were logged on 28th February and 1st March. This amounted to approximately five per cent of all calls throughout the year.
4. *Summative test window:* A large volume of calls was – unsurprisingly – related to the main test window. This volume of calls seems to have twin peaks; one just before the start of the window, and one right at the end of the window. This backs up other evidence that many schools ran test sessions towards the end of the four-week window.
5. *Results release*: There was a volume of calls relating to the release of results – almost all such calls related to how to access results, rather than questioning the actual results received.

165. Thus, an understanding of these potential causes of high call volumes, and the ongoing active drive to get schools to progress through accreditation, software installation and similar activities should help to increase the scalability of the solution for future years.

4.2.2.1.4 Service provision lessons learned

166. The RM service delivery report forefronted the following list of issues as the main lessons learned from the 2005 pilot:

- Schools require various levels of assistance to plan and implement successful test sessions, not all of which can be managed via training materials.
- Increased promotion of training materials and support articles is needed.
- Schools may experience internal issues, which impact their successful participation in the pilot. These are often beyond the control of the operational teams.
- Focus must be maintained on service levels; managing issues and enquiries raised by schools in a timely manner.
- Moving timescales or imposing deadlines has a negative impact on schools completing activities.
- Schools are willing to provide feedback on all areas of their participation.

167. Focus on these and other lessons learned should improve the scalability of the support services.

4.2.3 Evaluation of objective two

168. Thus, the overall high quality of support services, plus several other factors, suggests that the support provision should be scalable for a cohort of all secondary schools in England.

169. There are some issues that suggest a less hopeful scenario. In particular, there should be concern that several aspects of SLAs were not met. If the requirement for support were to increase in direct proportion with the number of schools to be involved in future pilots (65 – 100 per cent of schools to be involved in 2006, as opposed to roughly ten per cent in 2005), then it might be difficult to provide a scalable service.

170. However, there is also evidence that once a school has participated in a pilot, it seems to require less support for subsequent years. There is strong anecdote to support this from the experience of schools that took part both in 2004 and 2005, if not direct analysis. Thus, it would be reasonable to expect the schools that took part in 2005 to require less support subsequently.

171. But 2006 will be a year in which a major increase in participation will be required; as such scalability will be achievable only if support services are of the

highest quality. However, the preceding sections of this report give grounds to hope that such quality can be attained.

172. Therefore, objective two has been achieved. Once again, given a continued commitment to deal with current and future issues, it is likely that the infrastructure and support services can be suitably scalable to support the administration of a live National Curriculum test in 2008.

4.3 Objective three

173. 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.

174. 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.

175. 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 Previous findings on objective three

176. The interim report evaluated the formative reports only, since, at that time, the summative reports had not been delivered to schools.

177. The interim report found significant causes for concern with the formative reports and stated that they were likely to not achieve their objective.

4.3.2 Aspects of objective three evaluated in the final report

178. Given the aspects of the objective that were evaluated in the interim report, and the totality of the objective, it follows that the final report will:

- look for further evidence on the formative reports.
- fully evaluate the summative reports.

4.3.2.1 Formative reports

179. Although some further evidence that puts formative reports in a more positive light has become available since the writing of the interim report, it is not sufficient to claim that formative reports have achieved their objective.

180. It remains clear that most teachers and pupils did not see formative reports. The 33 teachers who did see the reports, and rated them in a questionnaire item, gave them a rating of 5.73 out of 10. This rating is an improvement on the rating of 3.13 that the formative reports were given at the time of interim evaluation. However, this improvement partly reflects the fact that teachers who claimed to

not have seen the formative reports and who therefore rated them as 'zero', were excluded from the analysis. Thus, it is not an adequate rating for the formative reports to achieve their objective.

181. Teachers commented that the statements were very similar, did not suggest areas in which pupils could improve, nor were they sufficiently specific.

4.3.2.2 Summative reports

182. In July 2005, QCA instructed RM to release results to schools without summative reports.

183. The purpose of summative reports was to:

- report individual pupil test results against the NC levels of attainment;
- identify pupils' overall level of attainment against school, local and national targets for Key Stage 3;
- give summary information about individual pupils' attainment for reporting to parents and, if appropriate, the next teacher.

184. Summative reports were designed to be based on the Key Characteristics of level. Key Characteristics are developed by QCA as an adjunct to the National Curriculum level descriptions. They are available on the NC in Action web site.

185. Summative reports were designed to be shorter and less detailed than formative reports. But they were intended to contain some information about the tasks that the pupil receiving the report undertook in the test, and a brief description of the things that that pupil showed s/he could do in the test. The emphasis in a summative report was to be on justifying and explaining the level obtained.

186. Summative reports were designed to give a maximum of four statements to justify pupils' levels.

187. However, prior to the release of summative test results on 12th July 2005, final reviews of the summative reports gave rise to the following concerns:

- Several level three, four, five and six pupils had a single statement. This was deemed not to provide sufficient information to justify the level of attainment;
- Two level six pupils had no statements at all. This was resolved by generating a manual report for these pupils.
- Upon reviewing reports that had three or more statements, QCA were still not satisfied that sufficient information with the required specificity to the pupils' performance was provided to justify the level of attainment.

188. Given these problems, summative reports were not released to schools.

4.3.2.3 Lessons to be learned

189. Following the non-release of the summative reports, RM and QCA agreed the following set of lessons to be learned on reporting:

- RM and QCA should consider changing the specifications for both 2006 formative and summative reports to facilitate the dissemination of more detailed information relating to pupil performance to schools.
- The design of 2006 summative and formative reports should be agreed early on within the 2006 test development cycle and samples should be produced to ensure consensus on what information they will contain and how this information will be set out.
- Test cases[4] for the generation of formative and summative reports need to be defined and agreed with QCA prior to testing of these reports commencing.
- Testing of the process for generating (and re-generating) 2006 formative and summative reports should be carried out well in advance of release decisions to ensure any issues can be resolved within the schedule. This could be achieved by using 2005 pupil data or 2006 pre-test data.
- To ensure quality, due diligence reviews of both formative and summative reports should be carried out by both RM and QCA prior to the release of practice and summative test materials. This could be achieved by using 2005 pupil data or 2006 pre-test data.

190. Additionally, one strand of the QCA formative evaluation will focus on the formative use of e-assessments.

4.3.3 Evaluation of objective three

191. Teachers' relatively low rating of formative reports, and the non-release of summative reports means that objective three has not been achieved.

[4] A test case has been defined as 'a method of exercising a product, feature, or process flow to confirm a single predefined result. Failed test cases reveal product defects or defects in the test case.'

4.4 Objective four

192. 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.

193. 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.

194. The reason for this change was that the security of the solution had been neither specified nor assessed prior to the pilot. Therefore, there were no grounds for assuming the solution to be secure.

195. The effect of this change in wording was to make the objective easier to achieve.

196. 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[5] 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.

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

198. 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 Previous findings on objective four

199. Evaluation of objective four in the interim report split the objective up into:

- Classroom issues
- Institutional issues

[5] 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.'

200. Classroom issues were defined as breaches of test security that occurred (substantially) in the classroom. Such breaches would include: pupils copying from each others' screens, and pupils or teachers learning about test content early in the test window and communicating that content to other pupils.

201. Institutional issues, by contrast, related to the disclosure of sensitive information (either confidential test content, or sensitive data relating to individuals or institutions) throughout the test development and delivery cycles.

202. Institutional issues could relate to the physical compromising of information (e.g. a briefcase left on a train) or the electronic loss (e.g. successful hacking of a server holding test data).

203. Institutional security issues could arise when a potential breach of security had occurred, as well as when an actual breach had occurred. For example, the leaving of sensitive test material on a train might necessitate the withdrawal of a test version, because there was a risk of unauthorised persons discovering test content inappropriately. This would be so, even if no unauthorised persons did in fact gain knowledge of confidential material.

204. The interim evaluation reported findings of trust management reports from the consultant, KPMG, and observations on the same topics from the Office of Government Commerce (OGC) Gateway review team.

205. KPMG's 2005 trust management reports suggested that the project had responded to 2004 reports (on arrangements for hosting test data on web servers, and susceptibility to external attack) by carrying out work that made the Key Stage ICT system less vulnerable.

206. However, 2005 trust management reports on new topics (security policies and procedures, and test development and management) suggested that substantial work was still needed to ensure security and confidentiality of information.

207. There was evidence of a high level of diligence within the project on security issues, and focused and pro-active management in this area from QCA.

208. In contrast to 'institutional issues' there was relatively little evidence on 'classroom breaches' of security. Small-scale (and therefore non-conclusive) observations of pre-test and summative sessions suggested that examination conditions were not being widely observed in the former case, but that they were being taken more seriously in the later test window.

4.4.2 Findings

4.4.2.1 Classroom issues

209. RM conducted a special trial to find out whether pupils could copy from each others' screens during a test session. In this trial, a test session was conducted under examination conditions (e.g. no talking) but pupils were encouraged to see if they could copy from other people's screens.

210. Unfortunately, several facets of the way in which the copying trial was implemented reduced the meaningfulness of findings from the trial. Such facets include:

- The fact that only one school was involved in the copying trial (reducing generalisability of findings).
- The fact that the copying aspect of the trial did not appear to be 'well advertised' to pupils (reducing the validity of the trial, since pupils may not have been trying to copy).
- The fact that only 24 sets of completed test results were returned to the central server from 62 pupils who initially took part (reducing the number of subjects for analysis of the potential benefits of copying).
- The fact that questionnaires used for sessions one and two appear to have had differently labelled options ('strongly agree … strongly disagree' for session one, 'yes .. no … don't know' for session two) may make it difficult to compare the extent to which pupils perceived copying to be possible in the two sessions.

211. The RM validity report advances some tentative findings on copying from the trial. However, given the limitations of that trial, these findings should not be considered to make a definitive statement about the issue of copying and test security.

212. As such, the test developer has not demonstrated that the test is secure from pupils gaining unfair advantage by copying. There are other grounds for believing that the test is secure (for example, the impact of cloning and randomisation making it unlikely for neighbouring pupils to get identical test versions – see interim report, paragraphs 214 – 216).

213. QCA has also recommended to the Programme Board that the security implications of test administration within schools should be investigated at the programme level.

214. However, the onus is on those who wish to introduce this test to demonstrate that it can be securely delivered (see paragraph 24 above).

4.4.2.2 Institutional issues

215. In addition to the trust management findings summarised above (paragraphs 204 to 207), it has emerged that there were in fact four 'institutional security breaches' during the 2005 pilot. These breaches are described in the following table:

Description of breach	Potential implications	Severity of breach	Actions taken
A member of RM content team sent an email without protection which referred to some elements of a modelling task.	The information could have been used to give some pupils unfair advantage in the described elements of the modelling task.	Minor	• RM staff reminded of security obligations. • No need to amend affected test materials.
A QCA staff member sent an unprotected email with overview of the content of the test.	Teachers in possession of overview could teach to the test.	Minor	• No need to amend affected test materials.
A draft teachers' guide contained a screen shot of the toolkit containing 2005 test material. The draft guide was used in the small-scale field trials of materials, and was sent to the printers, but was not distributed more widely, once the security breach was noticed.	Field trial schools and/or the printers might have distributed the secure content in the screen shot more widely. The breach would have been viewed negatively by all pilot schools if the guide had been widely distributed with the undetected secure information.	Minor	• RM staff reminded of security obligations. • List of RM staff having access to live test content reviewed.
Several problems on a single computer at Tata InfoTech were detected during routine monitoring. These included: • Poor firewall[6] configuration • Potentially ineffective anti-virus products in place • Development source code[7] potentially accessible via internet	A hostile attacker could potentially have maliciously introduced viruses, or damaged project source code. The absence of a formal security response process at RM hindered the speed of the investigation process.	Potentially high	• Tata took immediate action when alerted by RM. • Incident escalated to Tata senior management. • Machine in question was being used for study purposes only. • Source code on machine in question did not go back to the main repository[8].

Table 9: Institutional security breaches during the 2005 pilot

[6] A firewall is 'a dedicated computer or device with special security precautions on it, used to filter outside network, especially Internet, connections and dial-in lines.'

[7] The form in which a computer program is originally written, usually in a language which other programmers can understand. In order to actually run, the source code is changed by the computer's compiler into an internal language which is much harder for humans (but easier for the computer) to understand.

[8] A source code repository is a place where large amounts of source code are kept. They are often used by multi-developer projects to handle various versions and developers submitting various patches of code in an organized fashion.

216. The breaches described in the table varied in severity. Further, the project's response to the breaches would have been different if they had occurred with respect to the development of materials for a statutory National Curriculum test (breaches described as minor here would not have been thought so during a statutory administration).

217. However, given that these breaches occurred in a non-statutory development phase, they do not collectively suggest that the interim report finding that objective four had been passed should be overturned.

4.4.3 Evaluation of objective four

218. The findings section above confirms that the institutional breaches aspect of objective four has been achieved, whilst the classroom issues aspect has not.

4.5 Objective five

219. 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.

220. 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.

221. 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.

222. The presence of a second test window in 2005 does not necessarily imply that a second window will be run in future years.

4.5.1 Previous findings on objective five

223. The interim evaluation report findings with respect to:

- The accreditation phase
- Familiarisation materials
- The practice test
- The summative pre-test

224. The pilot was deemed to have achieved this objective (although there were a significant number of areas for further work).

225. A particular area of concern was that substantially fewer schools than the minimum that had been considered acceptable before the pilot actually participated.

226. The interim report did not consider the experience of schools in the second test window (due to the date of its authoring).

227. Given the facts stated in this section, the final report findings with respect to objective five will be concerned with schools' participation, and the experience of schools in the second test window.

4.5.2 Findings

4.5.2.1 Participation rates in the 2005 pilot

4.5.2.1.1 Numbers of schools ceasing to participate throughout the pilot

228. 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. (Interim evaluation report – paragraph 188)

229. In fact more than 45,000 pupils returned data, but only 402 schools took part. Thus, whilst there was more than enough data from pupils to provide information for validity work, participation rates of schools were lower than hoped. This has been a serious concern for the project and programme throughout the 2004 – 2005 cycle.

230. There were a number of steps that schools had to undertake before taking the summative test. The expression of interest, and the SMT commitment forms have already been mentioned (see paragraph 164 above). The schools also had to run the 'network audit' – a process to establish whether the schools' workstations and server were technically adequate to run KS3 ICT tests. As Table 10 shows, there were several ways in which schools could fail this audit. Following shipment of the software, there were two patches[9] to install – one to the Delivery Point System (DPS) and one to the Administration Point System (APS).

231. Table 10 illustrates the number of schools involved in the pilot at various stages (and, by implication, the numbers not proceeding to subsequent stages). The table is a snapshot of information available at 29th July 2005.

[9] A patch is a software update designed to repair known problems in previous software releases.

Activity	Number of schools involved in pilot
Expressed Interest	2447
Returned Forms	2051
Schools running Network Audit	1453
Passed Network Audit (more than 20 workstations)	1030
Passed Server but had fewer than 20 workstations	116
Failed either server or workstations	281
Failed server and workstations	26
Accredited	923
Shipped software	921
Start of day installed & registered version 8.5	713
Installed DPS v 8.6	654
Installed APS v 8.7	600

Table 10: Numbers of schools participating at different points in 2004 – 2005 cycle

232. Table 10 shows that substantial numbers of schools ceased to be involved at several stages of the 2004 – 2005 cycle.

4.5.2.1.2 Reasons for schools ceasing to participate

233. The reasons for schools ceasing to be involved were not immediately obvious. Indeed, it may well be difficult for the project to find out the motivations, and reasoning of non-participating schools; since, by definition, such schools are probably not especially likely to respond to questionnaires and similar requests for information.

234. Some analysis has been done on a small number of schools that curtailed involvement in the 2005 pilot after the accreditation phase, and that gave reasons for not wishing to remain in the pilot. This analysis has been done in two phases. Firstly, schools' reasons for quitting the pilot were grouped into a small number of relatively broad-brush categories:

Keyword	Total
Software	16
Time	14
Network	6
Staffing	6
Other	2
Software and network	1
Grand Total	**45**

Table 11: Reasons for schools not proceeding, grouped by key words

235. This categorisation was useful in that it confirmed schools could struggle to deal with a new and complex piece of software. It also showed how schools' time pressures could make it hard for them to participate in this new initiative, and that there were sometimes local issues with schools' networks that prevented them from taking part fully in the pilot.

236. However, the broad categories alone may not be the optimum way to describe schools' reasons for quitting the pilot. Often, schools gave several reasons for not wishing to further participate, or one reason would seem to conflate two issues (for example, a school not having enough time, because of a software or network issue). Therefore, a more detailed listing of schools' issues was undertaken by the evaluator. These were then regrouped into cognate categories. This second, more detailed listing is shown in the table below.

Grouped issues	Detailed issues	Number of schools	Sub-totals
Ceased involvement along the way	No time to do familiarisation and practice	8	**17**
	Negative experiences of pre-test	5	
	Did not realise how much preparation would be involved	2	
	More technical problems in summative test than in pre-test	1	
	Did first summative session, but not second	1	
Reservations about summative testing?	Will do practice tests only	10	**12**
	Test was too stressful for pupils	2	
Positive experiences	Would like to do it next year	4	**11**
	Generally positive experience	3	
	Positive experience of support	1	
	Disappointed to have to withdraw	1	
	Positive view of software	1	
	Would like to use it later in the term	1	
Timetabling	Timetabling clash with other tests/exams	7	**10**
	General timetabling problems	3	
Sundry technical issues	Tomcat problems	2	**7**
	Needed to deploy workstations manually	1	
	Problem printing off passwords	1	
	Problems with APS connectivity	1	
	Test caused school's server to go down.	1	
	Postmaster[10] problems	1	
Technicians' difficulties	Technicians had difficulties installing	3	**6**
	Technicians' time constraints prevented installation	3	
Idiosyncratic school factors	Building work in school	2	**6**
	Pressures within the school	1	
	Technical problems at their end	1	
	Local security settings caused problems	1	
	Network's resources allocated elsewhere	1	
Staffing issue	Staff illness	5	**6**
	Technician left the school	1	
ICT teachers told not to do it.	LEA advised that they would need a lot of time to run tests	1	**2**
	SMT told them to withdraw.	1	
Other	Do not teach year nine ICT	2	**11**
	School expressed dissatisfaction	2	
	Took too much time	2	
	Took too much work	1	
	No communications about updates	1	
	Roles not appropriate to schools	1	
	Technical issues take too long to resolve	1	
	Did not like the test interface	1	
Grand Total			**88**

Table 12: Detailed reasons for schools not proceeding with 2005 pilot

237. Several reservations about the analysis underlying this table should be expressed:

[10] 'Tomcat' and 'postmaster' are third-party database products which the KS3 ICT tests must access in order to run.

- This analysis comes from a group of 45 schools that expressed reasons for no longer wishing to be involved in the pilot after accreditation – it does not necessarily represent all schools that left the pilot at all stages.
- Analysis was based on reasons collected by RM service teams at the time of leaving; schools have not been re-contacted to confirm why they left.
- This analysis groups detailed reasons into what are believed to be intuitive broader categories; in the medium term, it would be useful for similar analyses to be informed by developed theories related to institutions' (non-)uptake of ICT innovations.

238. Notwithstanding the above reservations, several observations can be sustained from data in Table 12.

239. The largest group of reasons for schools not proceeding seems to relate to schools finding it difficult to sustain involvement throughout the many tasks needed to complete a summative test.

240. A second large group of schools had an underlying issue in that they were willing to take the Key Stage 3 software, but they preferred to use it in practice, not summative, mode. This could have several connotations: firstly, negatively, it might be that such schools had an intrinsic opposition to external summative testing. However, a more hopeful interpretation for the project might also be plausible; it might be that teachers found the software and pedagogic approach of the test to be positive, but they did not wish to commit to a summative test in 2005 for their own reasons. This issue is worthy of further investigation.

241. A third large group of reasons sees schools declining to take further part in the 2005 project, but nonetheless being optimistic about the test – saying that they would like to do the test next year, that they had had a generally positive experience of the test and/or support services. Such responses offer optimism that the project can increase participation in future years.

4.5.2.1.3 Current and future activity to increase participation

242. There has been a major new initiative within the Key Stage 3 ICT assessment programme to address the issue of schools' (non-)participation. This is known as the Participation Task Force (PTF).

243. The remit of the test development project – and by extension QCA – does not extend to preparing schools to take part in test pilots. However, it has also been agreed that 2005 practice with respect to encouraging schools to participate has been reactive. The aim of the PTF is to proactively drive forward uptake of this new innovation (the tests).

244. The position of the PTF within the programme and project governance and implementation is shown in the following figure:

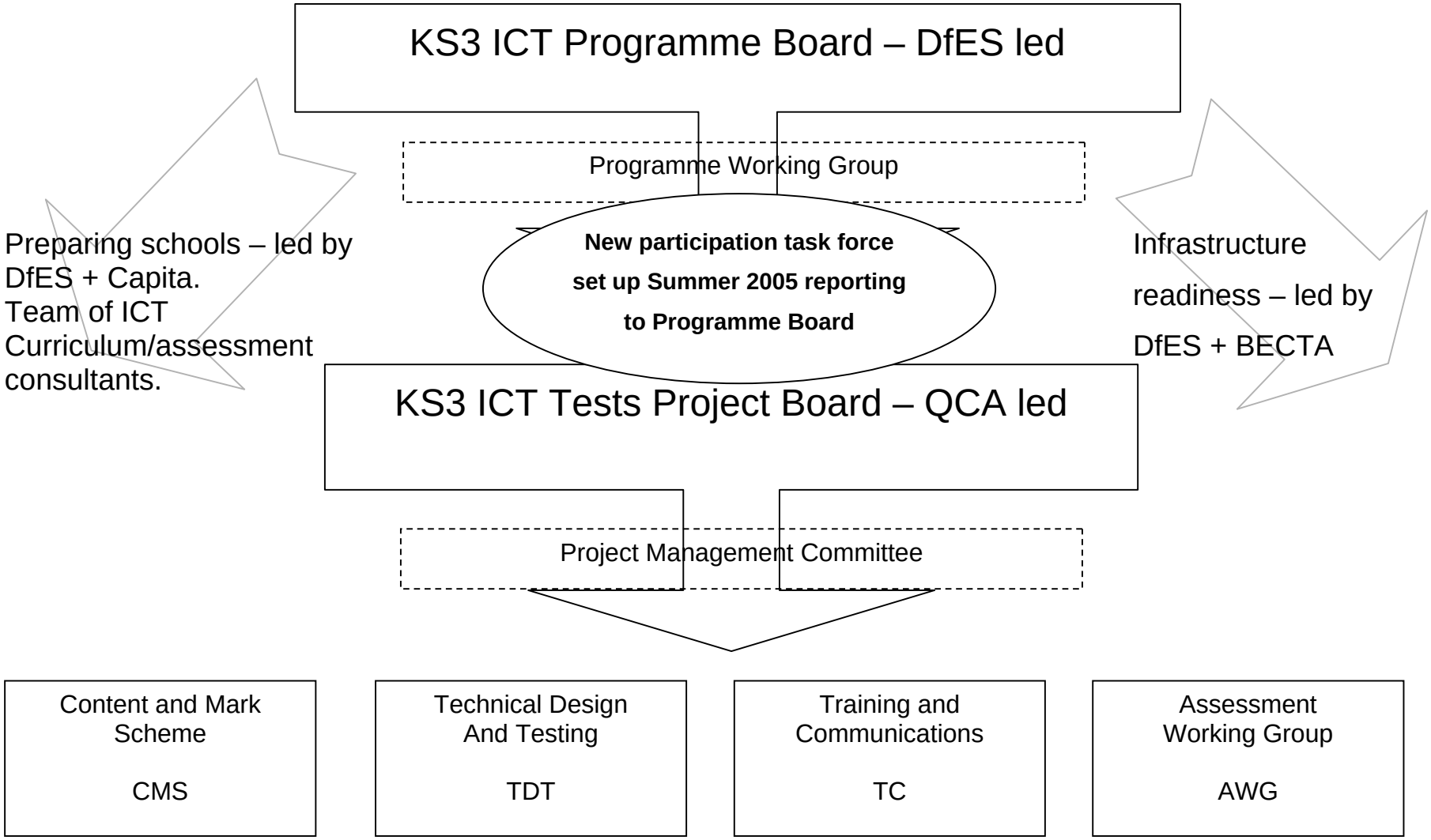

Figure 3: Place of Participation Task Force in relation to programme and project

245. The underlying purpose of the PTF is to ensure that all schools are prepared for the 2008 statutory test.

246. In support of this purpose the PTF has three main objectives:

- To achieve maximum participation for future pilots (60 per cent of secondary schools taking part in the 2006 pilot and 100 per cent running the 2007 pilot).
- To ensure successful experiences of the test by advising schools how to plan for the test effectively and aiding technical readiness.
- To understand and address the factors contributing to many schools' lack of involvement

247. It is believed that the structure described above will allow a wide ranging programme of work to be carried out by, amongst others: QCA and RM, the National Assessment Agency (NAA), and other strands of the Key Stage 3 ICT assessment programme.

4.5.2.2 Schools' experiences in the second test window

248. A summative test window was made available for schools that were not able to participate in the main, April 25th to May 20th window. This second test window ran from May 20th to July 15th.

249. 183 schools installed the software and were allocated test sessions, however, only 30 schools actually ran a full summative test in the non-pilot window.

250. No investigation has been run into the low uptake in the second window, nor has the usefulness of a second test window been evaluated.

4.5.3 Evaluation of objective five

251. Objective five was considered to have been achieved in the interim evaluation. Although the foregoing section gives rise to several areas for further investigation, there is nothing of major substance to revoke the finding that objective five has been achieved.

5 Evaluation in the light of overall project objectives

5.1 Purpose and approach of the test development project

252. The purpose of the Key Stage 3 ICT test development project includes the following:

> … to provide an independent measure of pupils' ICT attainment against National Curriculum attainment levels at Key Stage 3, in support of the new Public Service Agreement (PSA) target for ICT attainment at Key Stage 3. The test will initially complement teachers' own assessments of pupil attainment levels, which will be used to measure national progress against the PSA target. But over time, our aim is that the test will itself become the key progress measure, both nationally and for individual schools.

253. The benefits plan, which is appended to the 2005 pilot objectives includes the following primary benefits:

Tests
1. A valid, reliable and independent measure of Key Stage 3 ICT capabilities
2. A new innovative cutting edge model of testing

Teaching and learning
3. Improved understanding of standards in ICT and improved understanding of ICT capability
4. Improved pupil performance

Reporting
5. Formative feedback for pupils, teachers and parents
6. Detailed capture of data at all levels (pupil/school/LEA/National)

Administration
7. New innovative model for reduced bureaucracy in test administration
8. Automated test administration
9. Automated and consistent marking in moderation 'scripts'
10. No paper/speed of return of feedback

254. Additionally, DfES communications to QCA have emphasised that the project must deliver:

- a test capable of being put on a statutory footing by 2008
- a test that provides every pupil with an accurate National Curriculum level based on the KS3 ICT Programme of Study
- formative feedback on strengths and weaknesses

255. The project prefers to move towards full statutory implementation of the test in an incremental fashion, as outlined in the following statement:

> Our preferred approach is … to manage the risks of high stakes implementation by working towards that objective over a number of years, starting with a low stakes approach and moving through a planned process of trial, refinement and quality assurance.

5.2 Appropriate evaluation for a project in pilot phase

256. The Key Stage 3 ICT tests should be evaluated in a way that recognises not only their current pilot status, but also their potential high profile. As such, it is not required that the 2005 pilot tests are demonstrably perfect. However, stringent quality criteria are required.

257. In effect, in evaluating whether 2005 delivered a successful pilot, it was not expected that this year's pilot was of the same quality as a live National Curriculum administration. Rather, the best professional judgement of the evaluator will be applied to state whether it seems likely, given evidence available at the current time, that the tests will be able to be delivered in 2008 to the high quality that is required for National Curriculum tests.

258. In addressing this remit, the overall evaluation draws on the specific findings of previous sections of this report.

259. This overall judgement also makes reference to factors that are within the control of the test development project, and those which are not.

5.3 Findings

260. The 2005 pilot achieved the following objectives:

- Objective one: validity
- Objective two: scalability of infrastructure software and support services
- Objective four: institutional aspects of security
- Objective five: schools' experiences of the pilot

261. It failed the following (aspects of) objectives:

- Objective three: formative and summative reporting
- Objective four: classroom aspects of security

262. Whilst this is not a perfect result, it is a strong result for a project at this stage of its development.

263. There are many 'areas for further work' in the executive summary of this report. The evidence is that, since the publication of the interim report the project has worked on the 'areas for further work' from that document and used them to start to put in place remedial measures.

264. Thus, it is hopeful that the issues pointed out in this report can be further addressed.

265. Also, there are several developments beyond the formative evaluation that given hope that the Key Stage 3 ICT test can be realised as a high-quality product in 2008.

266. Firstly, a panel of external experts will investigate the 2005 level awards. Also, there are bespoke research activities into:

- Sources of difficulty in the test
- Formative reporting
- The strategy to base the test on a bespoke desktop environment

267. Furthermore, there is also evidence that the QCA Regulator and the Key Stage 3 project team will soon be working jointly to develop a regulatory environment that is suitable for this new type of test.

268. For the reasons advanced above, then, the overall objectives of the 2005 pilot have been achieved.

Annex A: Bibliography

American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) (1999) *Standards for Educational and Psychological Testing* Washington, DC: American Educational Research Association.

International panel (2002). *Maintaining GCE A level standards: the findings of an independent panel of experts*. [Online] available: http://www.internationalpanel.org.uk/ (October 3, 2005).

Royal-Dawson, L. (2005). *Is Teaching experience a necessary condition for markers of Key Stage 3 English?* Research paper prepared for QCA.

Wiliam, D. (2000). *Reliability, Validity and all that Jazz*. Education 3-13, **29** (3), pp. 9 – 13.

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
Sue Walton	QCA	Information relating to objective four, and general assistance relating to scope of evaluation, etc. Factual checking of draft report. General assistance relating to scope of evaluation, etc.
Hakan Redif	QCA	General assistance in providing information, and aiding delivery of the report.
Martin Adams	RM	Factual checking of draft report.
Mike Wright	RM	Responses relating to the amount of data that had been gathered in the summative pilot.
Louise Corder	RM	Factual checking of draft report.
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.
Gordon Nelson	RM	Factual checking of draft report.