Qualifications and Curriculum Authority

Evaluation of the 2006 pilot of the key stage 3 ICT test

Andrew Boyle
QCA Assessment Research team

September 2006

Contents

List of tables

List of figures

1 Executive summary

The Department for Education and Skills (DfES) initiated a programme for the assessment of information and communication technology (ICT) at key stage 3 (KS3).

The purposes of the KS3 ICT assessment programme are:

> to develop an on-screen test to provide an independent measure of ICT attainment at key stage 3 and to indicate the potential for the wider use of electronic testing in national curriculum (NC) tests and public examinations.

The Qualifications and Curriculum Authority (QCA) is running the KS3 ICT assessment programme on the DfES's behalf. QCA has engaged Research Machines PLC (RM) as its head contractor on the project.

During the 2006 pilot year, the test development project within the programme was transferred within QCA to the National Assessment Agency (NAA). Also, programme management responsibilities have transferred from the DfES to the NAA, following an Office of Government Commerce (OGC) review.

A pilot of a summative test was carried out in May 2006. 172,225 pupils in 1762 schools completed two sessions of the summative pilot test. These pupils' schools received a national curriculum level (between levels 3 and 6) and a summative report for each pupil who completed the test.

The 2006 test pilot was subject to seven objectives. This evaluation report considers the pilot's success with respect to the following objectives:

- Objective one: validity of the level 3 – 6 test
- Objective four: formative and summative reports
- Objective six: school participation and readiness

There were three possible findings for any objective:

- Objective achieved and demonstrated
- Achievement of objective not demonstrated
- Objective not achieved

Since the report does not evaluate all pilot objectives, no evaluation of the project's success against its overall purposes is made.

© 2006 QCA

However, the report concludes by offering a few observations about the general state of the KS3 ICT test. Firstly, the report notes that – although the three evaluated objectives were said to be 'not demonstrated' – the quality of information provided by project agencies for the evaluation was high, and consistent with a project of this importance.

Secondly, the concluding section points out several intrinsic tensions that are perceived within the test and its uses. These are listed below:

- The test is required to be both a novel e-assessment instrument, and to provide robust data to support summative accountability purposes. These dual purposes are a tension and a challenge to the project.
- The obligation remains upon the test development project to design a suitable test model and conduct necessary research to prove its validity.
- However, the fact that the test is required to be innovative means that a novel test model has been used. Correspondingly, those seeking to prove the test's validity have to use a range of novel and unproven methods. This can be seen as a positive development, but it is also a demand and a challenge.
- The evaluation of the KS3 ICT test has sought to support the test's development as a high-quality assessment suitable for full statutory roll out. However, it also sets a strict standard in making evaluative judgements; it will not judge a pilot objective to have been achieved unless there is strong evidence to that effect.

All the points above suggest that a degree of risk and uncertainty is associated with the test's likelihood of full roll out. Given that, it might be natural for decision makers to wish to 'de-risk' the project in some way – for example by requiring a more traditional test design to be produced.

The view of the evaluation is that the current dual strategy should be continued; that is, the development of a novel, difficult-to-prove test model, whilst requiring strong evidence of the test's quality before roll out. This is a difficult path, but – it is submitted – it is the right one.

Below are summaries of the report's findings with respect to each evaluated objective.

© 2006 QCA

Objective one: validity of the level 3 – 6 test

The achievement of objective one has not been demonstrated.

Validity was interpreted to consist of seven facets. Findings with respect of each facet are listed below:

Face validity

- Some commentators have questioned the decision to base the test on a bespoke virtual software environment. However, the evaluation finds the decision to use a bespoke virtual environment to have been correct.
- In the early part of the pilot cycle, there were perceived to be some issues affecting the quality of opinion collection and analysis across the project. However, QCA, NAA and RM staff have collaborated to take a set of actions to improve opinion collection and analysis. It is hoped that such actions will assure the quality of opinion gathering within the project.
- Teacher Review and National Stakeholder Groups found the test to be – on balance – face valid. However, both groups did express some concerns, or describe issues that they felt still outstanding.
- Relatively small groups of teachers and pupils returned questionnaires commenting on face validity issues. The findings from these questionnaires were mixed. There was a variety of evidence in both questionnaires supporting an argument that the test was face valid. However, responses to both instruments could also be interpreted to refute the contention that the test was face valid.
- The test development project intends to collect more opinion data relating to face validity in the autumn of 2006.

Content evidence of validity

- A teacher panel reviewed test content and found it to provide sufficient coverage across the ICT programme of study (PoS).
- There was some concern that this test could not assess working and communicating with others. However, it has been countered that such aspects of ICT can be considered to be assessed if remote (email) communication is included in the concept.

© 2006 QCA

Reliability

- Reliability is a fundamental property of effective measurement.
- Deriving reliability indices for an interactive e-assessment presents methodological challenges which are not present for conventional item-based tests.
- The key concepts in reliability analysis for the KS3 ICT test were test-retest method and classification consistency. In addition to analyses into classification consistency using the test-retest method, the project carried out back-up analysis using an approximation to conventional internal consistency methods.
- Internal consistency reliability indices for the 2006 summative test were shown to have improved from those derived in the 2005 test and the 2006 pre-test. But the increases in internal consistency reliability were quite small, and not universal across all levels of the test.
- Internal consistency indices for the 2006 KS3 ICT test summative pilot were compared with the most recent available indices for the existing NC tests at KS3. The indices for the existing tests were consistently higher than those for the ICT test.
- The argument can be made that the lower internal consistency indices for the ICT test reflected factors such as: the test's novelty or the less suitable data structure produced by the innovative interactive test. The low indices did not amount unequivocally to a finding of lower intrinsic reliability for the ICT test.
- The test-retest study showed 63 per cent of respondents being classified into the same level on two successive administrations.
- The project intends to do further research to derive a principled statistic to describe classification consistency. It will be very important to make sure that the 'experiment' that is set up to investigate classification consistency is carefully designed.
- This statistic will be derived for both the ICT test and existing NC tests. Well-designed analysis should facilitate a meaningful understanding of the classification consistency of the KS3 ICT test.

© 2006 QCA

Fairness for all pupils

- The best evidence currently available suggests that the test is fair for boys and girls and for pupils who are entitled to Free School Meals (FSM).
- Further work will be carried out to ascertain whether the test is fair for pupils who have Special Educational Needs (SEN), and who speak English as an Additional Language (EAL).

Concurrent evidence of validity

- Concurrent validity studies are often set up to have a sample of test takers sit a new test and an existing test of the same or a similar construct. This was not done in the current context, for reasons of practicality.
- Pupils' levels awarded by Teacher Assessment (TA) and by the KS3 ICT test were compared. 39 per cent of pupils were classified into the same level by TA and the test. The test classified six per cent of pupils into a level two or more levels below their TA level.
- Further work will be done in autumn 2006 to compare the consistency of classification between test and TA between the ICT test and existing NC tests. However, this evaluation has doubted the validity of such comparisons since TA and test may not be independent measures.

Construct evidence of validity

- A qualitative research project undertaken by the University of Leeds identified some sources of difficulty that might affect pupils' performance in the test.
- The test development project has responded positively to the Leeds findings and intends to use insights to improve the quality of test materials (for example, by giving fuller consideration to suggested sources of difficulty in the task writing and review processes).
- Further investigations in autumn 2006 will evaluate hypothesised correlations between aspects of ICT performance and aspects of performance on other NC tests. The results of these investigations will contribute to the argument about construct validity.

Level setting methods and outcomes

- An independent panel that reported on the 2005 level awards found that procedures to set levels were basically sound, although the levels derived in

© 2006 QCA

2005 could not be defended. Any weaknesses in the 2005 test that might have accounted for the 2005 level distribution were felt to be remediable before 2008.

- The independent panel suggested that the project team and contractor should have a reduced involvement in setting the educational standard. The project team contested this observation.
- There have been concerns that the summing of opportunities to represent 'increasing confidence' in a level award might not be defensible. The project has attempted to defend this concept as valid, however. NAA are organising a technical seminar in autumn 2006 to explore this further.
- Some concerns have been expressed about the conduct of the teacher panel to provide a set of cut scores to aid the determination of levels. Once again, these concerns are contested.
- 44.3 per cent of pupils achieved level 5 or above in the 2006 pilot.
- The distribution achieved in the 2006 pilot was some way above that achieved in the 2005 summative pilot, although still some way below the levels achieved in 2005 TA.
- There had been some commentary that TA in ICT had in previous years been too lenient. However, some ICT curriculum experts commented in August 2005 that any such leniency in TA was now less marked.
- It was noted that there was no obligation on the KS3 ICT test to equate to the TA distribution as if it were a pre-existing standard. Moreover, it was not possible to decide whether the TA or test distribution was the 'right one' because of the absence of an experimental design to support such a conclusion.

Overall evaluation of validity

- A broad definition of validity was agreed by project agencies, and a wide range of analysis was undertaken, and reported openly in well-written reports.
- The sub-objective with respect to 'content evidence of validity' was achieved, but for all other facets, validity was said to not have been demonstrated.
- Whilst there is a substantial body of planned research to address known, unresolved validity issues, there is no guarantee that such research will produce acceptable results.
- For these reasons, and some others, the evaluation is therefore that objective one has not been demonstrated.

© 2006 QCA

Objective four: formative and summative reports

The achievement of objective four has not been demonstrated.

Formative reports

- Formative reports were generated at the end of the second session of a practice test.
- Formative reports had the following features:
    - They were based on an individual pupil's performance in a practice test.
    - They were organised into the following sections: 'what you were asked to do', 'what you did', and 'how to make progress'.
    - They were available either on-screen, to save or print off, or via the school's Admin Point System (APS).
- A research project into formative reports was conducted in the first half of 2006. This project was based on three types of information: research literature into formative assessment, documents from the KS3 ICT test development project, and telephone interviews with teachers.
- The data from these three sources were not perfect (e.g. some literature on formative e-assessment was not well designed, and there were only a few telephone interviews). However, the research provided an appropriate information basis to generate and exemplify concepts relevant to the formative reports.
- Some key findings from the respect fields included:
- Formative assessment literature
    - Formative assessment could improve pupils' attainment.
    - There was a poverty of practice in formative assessment.
    - Feedback to a pupil should be about the particular properties of his or her work, with advice on what he or she can do to improve.
    - Pupils should be trained in self-assessment.
    - Formative assessment should be designed into any piece of teaching.
    - Written feedback should consist of comments, rather than marks or grades.
    - Researchers and test developers have used many different types of instrument to provide e-formative assessment.
    - Feedback from e-assessments has been delivered in many different ways and at many different junctures (e.g. after individual questions, at the end of entire test session, etc.).

© 2006 QCA

  - Researchers have claimed that formative e-assessment can have benefits not present in 'pencil-and-paper' formative assessment (e.g. the ability to tailor feedback for different learning styles). However, some have countered that e-feedback can lead to superficial behaviour (e.g. clicking through web sites) rather than 'deep learning'.

- Project documents
  - Key benefit BS001 of the KS3 ICT programme is:

    Immediate formative feedback from the practice test aids teaching and learning.

  - Original project documents envisaged the production of pupil-level and group-level formative reports. In principle group-level formative reports could be produced by current systems, but this would require further development work.
  - The project carried out substantial work between the 2005 and 2006 pilots to improve the formative reports.
  - Prior to releasing the 2006 practice test, credible software testing suggested that the formative reports were accurate and logical.
- Telephone interviews
  - Only a small number of teachers took part in interviews, despite a large number being approached.
  - Consistent findings amongst teachers who did respond included:
    - Formative reports should contain a national curriculum level.
    - There was some criticism of the language of the reports: that the language was aimed at teachers and not pupils, and that the reading load was too great for a year 9 pupil.
    - The reports were thought to be targeted at the wrong audience – teachers rather than pupils. There was a contrary strand of opinion, however. Some teachers thought that teachers should mediate feedback to children, and that therefore the audience for these reports was valid.
    - There should be formative reports designed for specific audiences: one for pupils, and two for teachers – one for an individual pupil and one for a group.

- Despite the improvements to the formative reports and the pre-release testing, the evaluation found that the lack of use of the reports meant that they could not

© 2006 QCA

be described as 'useful', and that therefore the objective had not been demonstrated.

- The formative reports research recommended the following:
    - The investigation of the development of a wider bank of formative assessment materials.
    - The production of three types of reports – two for teachers and one aimed at pupils.
    - National curriculum levels should not be added to formative reports.

Summative reports

- Summative reports were tested prior to their release, and the successful testing was considered to guarantee the accuracy of the reports.
- No evidence is yet presented about the usefulness of the reports – since they were only sent to schools towards the end of the summer term.

Objective six: school participation and readiness

The achievement of objective six has not been demonstrated.

Definition of eligible schools

- Deciding on how certain categories of schools should be treated for participation purposes is surprisingly complex. In particular, the way in which small non-standard schools (such as special schools, Pupil Referral Units – PRUs, hospital schools, etc.) are treated is an important consideration that affects participation statistics.

2006 participation

- The KS3 ICT assessment programme has categorised schools according to their participation status. This status has served as a framework for further work – such as the reporting of participation statistics and the deployment of best-placed agencies to encourage schools to participate.
- The 2006 objective and Critical Success Factor (CSF) for school participation were intentionally written to be very tough. They were also defined before the instigation of the NAA Wider School Readiness project (see below).
- The numbers of schools participating in the 2006 pilot fell well below the pilot objectives; this was true in respect of accreditation, installation of assessment system software infrastructure, installation of specific test software packages and actual running of summative tests.

© 2006 QCA

Schools' experience of the 2006 pilot

- A set of questionnaires administered with school staff at different stages of the 2006 pilot showed consistently high satisfaction ratings. This was particularly the case for early stages of the process (accreditation and software infrastructure) Success ratings were also high for contact with RM customer services. A small sample of questionnaires from the pilot window showed respondents still happy overall with their experience, but a little less so than for earlier stages of the process.
- The project operations team has compiled questionnaire responses to set out a credible set of lessons learned from the pilot.
- Reasons given by a set of schools who contacted the project to explain their non-participation in 2006 showed the biggest group that dropped out could be said to be 'waiting and seeing' what would happen to the KS3 ICT test, rather than proactively engaging. From this sample of school opinion there seemed less criticism of the test – or related issues such as administrative burden – than had been the case in 2005.
- Secondary National Strategy (SNS) consultants contacted non-participating schools on behalf of NAA. Schools' reasons for not participating were grouped into the main categories of: establishment, leadership, ICT staff, technical and pupils. This categorisation has been used to further organise encouragement to schools' participation.
- Particular findings from this SNS data collection included:
    - Small non-standard schools made up a large proportion of those who did not participate.
    - 'Fragile establishments' (e.g. those subject to special measures) also appeared less likely to participate.
    - ICT staff shortage was associated with non-participation.
    - Some ICT departments had not been participating because they 'had other priorities'.
    - There was alleged to be a substantial group of schools in which Senior Leadership teams (SLTs) had 'discouraged' participation.
- Whilst schools' experiences had been positive across the piece, they were not universally at the very high standard that had been set in the pilot objective. Therefore, this facet of the objective was not met.

© 2006 QCA

Plans to ensure full participation in 2007

- The KS3 ICT programme board has noted at least five risks which relate to the potential for less than 100 per cent participation prior to 2008. This shows that the most authoritative decision-making body concerned with the programme has the issue firmly on its radar, that substantial activity is already underway to promote participation and that substantial new pieces of work are planned to further promote the test.
- A project for Wider School Readiness (WSR) has been instigated. This project intends to fill a gap that previously existed in the programme; it addresses a wide range of non-technical administrative tasks that need to be carried out to successfully conduct pilots of the test.
- The WSR project has three main outputs: on-paper and online versions of an essential guide (EG), and increased support for NAA field support officers (FSOs) – who, in turn – support school Exams Officers (EOs).
- Early indications of progress towards 2007 participation targets show participation rates slipping behind targets.
- Whilst the wide range of activity with respect of participation and wider school readiness gives good grounds for believing that 100 per cent participation will be achieved in 2007, the missing of early targets – added to the experience of past years missing participation targets – leads to the evaluation that a credible plan for 2007 participation has not yet been demonstrated.

© 2006 QCA

2 Introduction

2.1 Purpose of this report

1. This report evaluates the 2006 pilot of the KS3 ICT onscreen test.
2. The report fulfils the following main purposes:
    - It provides an evaluation of the status of the key stage 3 ICT test for and on behalf of the KS3 ICT project team. The report will also be sent to members of the KS3 ICT Programme Board.
    - It provides information for the QCA Regulations and Standards division to inform their subsequent audit of the KS3 ICT test as a step in the statutory readiness process.
    - It provides evidence that the KS3 ICT test project and relevant aspects of the KS3 ICT assessment programme are actively and thoroughly evaluated.

2.2 Organisation of this report

3. Background information on the key stage 3 ICT assessment programme and the test instruments that were designed can be found at pages 85 and 88, respectively.
4. At the start of the report proper, there are short 'Method' and 'Evidence' sections (pp. 13 and 18).
5. The main findings of the report start on page 19. The findings are organised by objective. Objective one (validity) starts at page 19. Objective four (formative and summative reports) starts on page 52. Objective six (schools' participation and readiness) starts on page 59.
6. At the end of the report there is a short section setting out some concluding remarks (page 82).

© 2006 QCA

3 Method

3.1 Evaluation across the project

7. The key stage 3 ICT test development project reviews progress towards objectives through a wide variety of means. These include:

   - The conduct of several management group meetings, at which details of forthcoming work are agreed and progress against existing targets is monitored.
   - The conduct of project (until February 2006) and programme boards, at which issues from the management groups and elsewhere are fed up to suitable forums to allow decisions to be made.
   - The maintenance of logs from management groups and the boards to record issues, risks and so forth, which could affect the project.
   - The involvement of stakeholders through a variety of means, such as: Teacher Review Groups, National Stakeholder Groups, RM Teacher Panels, etc.
   - The following of clear procedures by QCA/NAA to sign off material before it is sent out to schools (for example via acceptance testing and approval of RM release recommendations).
   - The active gathering of opinions, for example via questionnaire surveys.
   - The analysis of pilot outcomes in detailed reports created by RM and their sub-contractors; these reports being made available to NAA and QCA staff at important meetings (e.g. level setting) and otherwise.
   - The addressing of specific issues of concern to the project by the running of freestanding, bespoke research activities (for example to review the sources of difficulty affecting pupils taking the test and the review of the formative reports within the test).

3.2 Formative evaluation

8. One specific area of activity that supports the project is a formative evaluation. This work provides information to assist the project in improving its products on an ongoing basis; commenting on work whilst it is being done, rather than waiting until it is completed.
9. The formative evaluation produces a range of products; some formal and some informal. Such evaluative inputs are transmitted through some of the channels referred to above – for example as content in meetings, or as reports on bespoke research projects.

© 2006 QCA

3.3 Summative evaluation

3.3.1 Definition

10. The current report is a summary of work done during a cycle of the project, presented along with judgements as to the work's successfulness, to external readers. It is, therefore, summative evaluation.
11. This report has been commissioned by the test development project, but is written by a researcher from the QCA Assessment Research team. It is presented by that researcher in his own right, but also on behalf of the project team. This arrangement is different from many summative evaluations, which would generally be conducted by a wholly independent third-party.
12. The arrangement is justified on the grounds that the writing of the report is substantially independent, and that any content changes that have been required by the project are negotiated by the project and the evaluator in accordance with a set of principles that were set out before the writing of the report (see para 25 below).

3.3.2 Objectives

13. The 2006 pilot is subject to a set of objectives, and critical and other success factors devised by the programme board (FINAL v0.8 – 21 December 2005). These objectives and success factors are similar to those used in previous years, but they have been defined to make it more straightforward to objectively evaluate.
14. This report will cover a selection of the 2006 pilot objectives. Those are:
    - Objective 1: Develop and administer KS3 ICT tests that deliver a valid and reliable assessment of pupil performance and award defensible national curriculum levels 3 – 6.
    - Objective 4: Provide all schools participating in the 2006 pilot with accurate formative reports from the practice test and an accurate summative report from the summative tests.
    - Objective 6: Ensure that schools register interest in the pilot, complete accreditation, move to Approved Test Centre status and with adequate preparation time, take part in the 2006 summative test window, and have a satisfactory experience throughout the process.
15. This set of objectives excludes those to do with software and support services robustness, and trust management. This reflects the experience from 2005 that it could be tricky for the evaluator to make judgements about issues which were beyond his professional expertise. It is also consistent with the approach

© 2006 QCA

currently within the programme, to instigate a set of 'statutory readiness criteria' and for each criterion to be 'owned' by a group of experts in the specific field.

16. Objective 2 – which relates to the level 7-8 test that was trialled in 2006 – is also not evaluated in this report. This is due to it not being clear, at the time of writing, what lessons have been learned from that trial.

3.3.3 Overall evaluation of the pilot

17. Because not all objectives are addressed in this report, it will not be possible to evaluate the project's overall status with respect to the totality of its purpose. Rather, evaluation judgements will be confined to the objectives under scrutiny.

3.3.4 Evaluation of the project, not the contractor

18. Formative and summative evaluation aims to provide useful information to allow the project to improve its products. It does not directly comment upon the activities of the contractor carrying out work – Research Machines, in this case. Of course, the contractor may need to take actions as a consequence of evaluation findings, but it is important that the focus of the evaluation is noted – on the project, not the contractor.
19. This focus affects the terminology used in this report. Often actions of 'the project' are referred to. This suggests a degree of joint authorship and agency; the test development agency will write a specification document or a report, staff at QCA/NAA will advise on the content of the report, require changes and accept the report when they believe it to be of suitable quality. In that way the document will become 'the project's' document.

3.3.5 Interaction of evaluation and regulation

20. The 2006 test was a pilot; it was not run for high-stakes purposes. It follows that it should not be judged as though it were used for high-stakes purposes.
21. During 2006 QCA's national curriculum assessment monitoring staff from Regulation and Standards Division have been activity engaged with the KS3 ICT test. This engagement has taken several forms:
    - Mapping findings from the 2005 evaluation report against regulatory framework common criteria
    - Observing level-setting meetings
    - Receiving and commenting upon project specification documents
    - Being interviewed for this report.

© 2006 QCA

22. In 2006/07 and thereafter, it is intended that the Regulator will be involved in work on the 'statutory readiness criteria' (see para 15 above); thus ensuring that the project benefits from joined-up thinking in the run up to full roll out.
23. Despite this engagement, however, the 2006 pilot was not a regulated test. It was not subject to a full audit, as would be the case with a statutory test.
24. The current report will be copied to the Regulator, to help them to formulate a set of issues that need to be considered for the regulation of this novel test.

3.3.6 Interpretive approaches used in evaluation

25. Any evaluation requires a professional judgement. Such a judgement can be difficult to define objectively in advance. However, in order to maximise transparency, interpretive approaches that will assist the arrival at evaluative judgements have been communicated to the project via the Assessment Working Group.
26. This is a summary of some interpretive approaches:
    - The evaluation will fulfil the mission of the QCA by putting the interests of the learner first.
    - There is an active obligation on the project to demonstrate that the test has satisfied its objectives. Whilst any statements that the test has not achieved its objective will be supported by evidence, the substantial burden of proof that the project will have to discharge for an objective to be satisfied should be noted.
    - It follows from the above that there will be three possible findings with respect to objectives:
        - Objective achieved and demonstrated
        - Achievement of objective not demonstrated
        - Objective not achieved
    - The standard to be achieved in evaluation will be the same as in previous years. However, since the 2006 pilot is one year closer to full roll out than 2005 was, the phrase 'this test was valid, *given that it was a pilot year*' will be used more sparingly than previously.
27. This report will provide evaluation in specific areas (for example: validity and formative reports). Relevant definitions will be provided at the start of each substantive section.

3.3.7 Quality assurance and ethical controls

28. The bases of this report were defined in a Product Description before its inception.
29. In writing this report the evaluator is governed by the QCA's research code of ethics.

© 2006 QCA

30. Several checking stages are incorporated in the production of this report. They include:

- Vetting by senior colleagues to ensure that methods are in line with best practice, and that findings are defensible and appropriately expressed.
- Vetting by the project director to negotiate an agreed content (see also para 11 above).
- RM has provided a factual check (they are not invited to comment on issues of interpretation).
- The revised report is proof read by a researcher.

© 2006 QCA

4 Evidence

31. This report makes use of the maximum range of available reliable information. Such information includes:

    - Documents (such as specifications and reports) provided by the project contractor – RM.
    - Minutes of meetings, project logs, and other project documentation.
    - Reports of other research projects conducted throughout the year. Such projects include:
        - The report of the independent panel that reviewed the distribution of levels awarded in the 2005 pilot
        - Documents suggesting the most appropriate models to use for reliability analysis and reporting
        - An information paper prepared for the QCA Executive on the decision to base the KS3 ICT tests on a bespoke virtual software environment.
    - Information from informal advice given throughout the year (for example with respect to the collection and analysis of opinions within the project).
    - Interviews with project staff and other key stakeholders which were specially conducted for this project.

32. The major source of evidence for validation work in the 2006 pilot was that generated by pupils sitting the test. 172,225 pupils in 1762 schools completed two sessions of the summative pilot test. These pupils' schools received a national curriculum level for each test taker.

© 2006 QCA

5 Findings

5.1 Objective one

33. Objective one is:

> Develop and administer key stage 3 ICT tests that deliver a valid and reliable assessment of pupil performance and award defensible national curriculum levels 3 – 6.

34. The Critical Success Factor associated with objective one is:

> This CSF is met if the 2006 'validity and reliability' 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 which accurately awards Levels 3 – 6 or N to pupils completing the test.

35. The other success factors are:

- All pupils completing the levels 3-6 test in the 2006 pilot are awarded and receive a national curriculum level or N and supporting summative report.
- Appropriate analysis, using an appropriate range of methods, and reporting on experiences of and feedback from appropriately sized samples of schools and stakeholders, demonstrates to QCA and DfES's approval, that the test is a valid assessment of the ICT Programme of Study.
- Appropriate analysis of validity is conducted. Validity reports set out transparent and defensible methodologies for interpreting that analysis.
- QCA and DfES accept validity statements.
- Level awarding procedures are in accordance with their specification and best practice in the field.

5.1.1 Definitional issues

36. Validity has been widely agreed to be the central concept in understanding the quality and appropriateness of a test and its uses. It has had many definitions; however, in the current context it has not been appropriate to adopt a single definition of the concept wholesale.
37. Rather, it is easier to understand the practical import of validity for this test development by examining several of its features. Whilst this does not provide an incontrovertible explanation of validity, it may allow the reader to appreciate what validity and its investigation meant in the 2006 pilot of the KS3 ICT test.
38. Firstly, the stringent standards that are placed upon the test development project should be noted (see para 26 above). This strict set of interpretative criteria means that the test development project is under an active duty to provide clear evidence that the test is valid.
39. Further, validity was taken to be made up of several facets. These included:

© 2006 QCA

- Face validity
- Content evidence of validity
- Reliability
- Fairness for all pupils
- Concurrent evidence of validity
- Construct evidence of validity
- Level setting procedures and process

40. Each facet of validity will be briefly defined at the start of its sub-section.
41. A further issue that has concerned the evaluation and the project more widely is whether validity can be taken as a single unitary concept with a number of facets, or whether it is better to conceptualise a group of distinct validities, or types of validity.
42. The evaluation has adopted the position that validity should be viewed as a single, indivisible construct. This is for the following reasons:
    - To allow a single evaluative judgement to be made as to whether the 2006 pilot test was sufficiently valid or not.
    - To emphasise that all facets of validity are necessary conditions – for example to negate any tendency to promote a particular facet of validity as *prima inter pares*.
43. Thus, a 'micro-evaluation' of each facet of validity will be made at the end of each sub-section. The overall evaluation of validity will be informed by referring to all the micro-evaluations of the validity facets.

5.1.2 Face validity

44. The validity specification document defines face validity in the following terms:

    > Face validity seeks to provide evidence that stakeholders find the test instrument and outcomes meaningful and accurate. An additional benefit would be that the test experience is positive for pupils and staff but this is not in itself necessary for the test to show face validity. It is concerned with whether or not a test, and its constituent items or mark points, seems to measure what it is claimed to measure.
    >
    > A test with high face validity may have a better chance than an equivalent test with low face validity of inducing co-operation and positive motivation among subjects before and during the test administration, reducing dissatisfaction and feelings of injustice among low scorers, convincing stakeholder groups such as policymakers, users and administrators that the test is trustworthy thereby generating positive public and media relations.
    >
    > It is important that while a test may be technically valid and reliable, it is also perceived as appropriate by the users.

45. The validity report describes findings from teacher review and national stakeholder groups. In addition to summarising such findings, this sub-section of

© 2006 QCA

the evaluation report addresses issues concerning the use of a bespoke virtual software environment to support this test, and opinion collection and analysis.

46. The opinion collection findings are included under 'face validity' because they impact on the extent that ways in which one should interpret reports of stakeholders' opinions about the test. The issue of the bespoke virtual software environment is one that teachers and others frequently mention when commenting on the test. Thus, it is also included in the face validity section.

5.1.2.1 The bespoke virtual software environment

47. The KS3 ICT test is based on a bespoke virtual software environment. The environment is bespoke in that it is designed especially for this test, and – whilst it has many features in common with other software suites – it is not identical to any proprietary suite of applications. The applications are virtual in the sense that they cannot create outputs (e.g. files; emails; documents to be printed) that can be transferred to a different, 'real world' software environment (see paras 384ff for a more detailed description of the bespoke environment).
48. There has been some criticism of the reliance on a virtual bespoke environment (see para 69 below, which lists teacher questionnaire findings). The criticism has come from two directions: some commentators have queried whether the test should be based on a widely-used operating system and software suite (for example, Microsoft Windows and Office). A separate strand of comment has suggested that QCA/NAA should develop the existing KS3 ICT programs more fully – so that they will be a fully functioning suite of office-type programs.
49. It is the view of this evaluation (based on study of previous documents and interviews with several relevant members of staff at QCA/NAA) that the current decision to base the test on a virtual suite of bespoke programs is the correct one. It therefore follows that no move should be made to base the test on a proprietary operating system and suite of programs, nor should the existing programs be worked up to make them a fully functional suite of applications.
50. The main reasons in favour of the bespoke virtual environment are:
    - Several pilots have been conducted and demonstrate that the virtual office suite can form the basis for valid measurement.
    - The current suite of virtual programs is central to the test that QCA has been developing over several years; it would not be possible to 'port' the assessment model over to a different suite of applications. Thus, any decision to change this fundamental tenet of the test would risk losing several years' worth of work.
    - By developing an entirely bespoke solution, QCA has complete ownership of Intellectual Property Rights (IPR).

© 2006 QCA

51. The strongest reasons *against* basing the test around a propriety suite of software programs (such as Microsoft Office) include:
    - Allowing pupils to use functioning 'office-type' applications would also allow them to use the internet; as such, it would be impossible to guarantee secure test conditions.
    - It would be very difficult for test developers to provide fair measurement for users of the many different versions of existing software applications.
    - It has not been demonstrated that the necessary technology exists to allow a test to track and capture pupils' actions in order to compare them with a measurement model.

52. The strongest reasons *against* the further development of the KS3 virtual toolkit to make it into a fully functioning software suite include:
    - It is not part of QCA's business to develop nor to provide long-term support for functioning suites of 'office-type' applications.
    - QCA should not attempt a substantial piece of development without a clear demand for it from the DfES.
    - Any suite of functioning 'office programs' that QCA could develop would inevitably be much less powerful than existing commercial suites.
    - The current focus of QCA's work is to provide a valid test that can be taken up by all schools in England. This is proving to be a hard enough task in itself. Any other development would be a distraction from this central task.

53. This advice has been communicated to the QCA Executive.
54. Despite the clear finding that the current solution of a bespoke virtual environment is the correct one, there are arguments that can be made in favour of either basing the test on a proprietary solution or of further developing the bespoke solution. It is likely that some observers will continue to make such arguments. It is thus important that the project – and QCA more generally – should defend the decision to base the test on a virtual desktop environment.
55. However, there are two senses in which the advice given to the QCA Executive was not complete. Firstly, the project and the formative evaluator's advice was mainly based on opinion of QCA staff members. Secondly, the project was not able to provide clear statistical evidence from a valid and reliable external measure (or combination of measures) that pupils who were being awarded a given level by the KS3 ICT test would be likely to be awarded the same level by the credible external measure.
56. The first omission in the advice to the Executive will be rectified by NAA's commissioning of an independent consultant to review the decision to use a bespoke virtual environment.

© 2006 QCA

57. There are a number of issues surrounding the production of concurrent evidence of validity. These will be outlined in the relevant sub-section of this report (see paras 143ff).

5.1.2.2 Opinion collection and analysis

58. The KS3 ICT test development project makes substantial efforts to discover the opinions of stakeholders. It seeks the views of national stakeholders (e.g. British Educational Communications and Technology Agency (Becta), Office for Standards in Education (Ofsted), the Secondary National Strategy (SNS), DfES, etc.), local authorities, teachers, other school staff, pupils, and so on. It runs a range of opinion-collection exercises – for example using instruments such as questionnaires, case studies and focus groups.
59. In 2005 the project was praised by the OGC for its opinion-collection work.
60. A piece of formative evaluation work was undertaken collaboratively by the evaluator, the NAA project director and RM project staff. This work addressed the suggestion from the evaluator that:
    - It would be useful to construct a matrix showing and categorising all the opinion collection that was being carried out across the project. This could then be used to make sure that there was a suitable balance of opinion collection in the project.
    - It was important that questionnaires were rigorously proof read before being sent out to users (for instance to make sure that they complied with a coherent house style and that they did not contain literal typos).
    - It was important to make sure that opinion-collection instruments were not unduly positive in orientation (for example if questionnaires routinely required respondents to agree or disagree with a positive statement about part of the project this might introduce an unrepresentatively positive representation of stakeholder opinion).

61. The collaboration on opinion collection and analysis resulted in the following actions:
    - A matrix of opinion collection and analyses has been created by the project. This matrix shows all opinion collection and analysis across the project, and thus allows a reviewer to make sure that in any particular area of the project suitable amounts and types of opinion collection are being conducted.
    - In future all questionnaires will be reviewed by specifically relevant project teams.
    - Questionnaires that have been written since the start of this collaboration have contained questions written in a range of styles – not just positive statements with which the respondent was invited to agree or disagree.
    - A piece of work has been carried out by RM to improve the quality of evidence gathered from school visits. This work involved:
        - Revising the instruments used to standardise the collection of data on school visits.
        - Conducting training sessions prior to staff visiting schools.

© 2006 QCA

- Evaluating the efficacy of the school visits pro-forma, after it had been used in visits.

62. It is hoped that these actions will lead to more valid and reliable opinion collection and analysis.

5.1.2.3 Stakeholder review groups

63. The project conducts a range of activities in which stakeholders are invited to express their opinions on the test. These include a Teacher Review Group (TRG) and a National Stakeholder Group (NSG).
64. These groups were conducted throughout the 2005-06 development cycle and focused on specific issues within the project. They were run and minuted as meetings.
65. The project validity report described findings from a TRG held on 30th January 2006. That group commented on the level 4 – 6 tier of the test.
66. 17 positive comments on the test are recorded in the report. These include:

- Tasks are excellent and test pupils' ICT competencies very well.
- Vast improvement from last year – interesting and challenging tasks – well done!
- Liked this – a big step forward and really does need pupils thinking about how to use ICT rather than just knowing skills. Definitely looked at application of ICT. It will sort out L6 from the lower levels.

67. The report also listed 15 'issues'. It suggested that three of these were key:

- There is still an issue with the assessment of reviewing, modifying and evaluating work in the 2006 test.
- Some of the functionality in the toolkit needs to be improved or amended.
- Pupil familiarisation will be vital.

68. The NSG also reviewed the level 4 – 6 tier of the test, but on 27th January 2006. It is reported that this group was happy overall with the test, and with changes made as a result of previous reviews. The NSG's only reported concern related to pupils' use of the style guide in the final task.

5.1.2.4 Questionnaire findings

5.1.2.4.1 Teachers

69. The responses of 68 teachers to a questionnaire are analysed in the validity report. That report describes the following main findings:

© 2006 QCA

Statement	Number of respondents
The tasks were appropriate for pupils.	38
The task instructions were clear.	40
The test was a suitable way of assessing pupils ICT capacity.	34
The test was of appropriate difficulty.	33
The test covered an appropriate range of the KS3 ICT programme of study.	44
50 minutes was of a suitable duration for a test session.	48
The workstation software was reliable.	34
The word processor was intuitive for pupils to use.	43
The spreadsheet was intuitive for pupils.	37
The database was intuitive for pupils.	48

Table 1: Summary of a sample of 68 teachers' views

70. In 'free-text comments' 25 teachers stated that the software should be more like Microsoft applications and Windows operating system.

5.1.2.4.2 Pupils

71. The responses of 434 pupils from 16 schools are recorded in the validity report. Key findings from the pupils' questionnaire responses include:

- Over half of the pupils were able to understand the task instructions (58%).
- Over half of the pupils did not have enough time to complete the test (52%).
- Over half of the pupils were able to show their ICT ability in the test (51%).
- One third of the pupils used pencil and paper in the test (33%).
- Approximately two-thirds of the pupils found the test difficult or very difficult (69%). The main reasons cited for the cause of this difficulty were:
  - 'test software could not do what I wanted it to do' (23%),
  - 'not understanding the task instructions' (20%) and
  - 'not understanding how to do the tasks' (18%).

72. The top three main suggestions for improving the test were:

- More time or less to do in the test
- Clearer and simpler instructions
- Make the software more like Microsoft or generic software

73. Most pupils rated word processor and presentation as easy to use. Only about half of the pupils rated database, spreadsheet and web browser easy to use.
74. Pupils liked using the computer for the test, emailing and learning and experimenting with the software. Many pupils commented that there was 'nothing' they liked about the test or that they liked 'finishing the test'.

© 2006 QCA

5.1.2.5 Evaluation of face validity

75. There are substantial grounds for believing that the test is 'face valid'. However, there is also some evidence to question such claims.
76. The view of the project, in its validity report, is that the test has 'the potential to be face valid'. To further investigate this area, a short, targeted questionnaire on the face validity of the test, and the levels awarded will be sent to school ICT co-ordinators in early autumn 2006.
77. The further investigation into face validity is imminent, and can be welcomed. This work should allow the project to understand schools' perceptions of the test, and allow changes to be made, or information to be better communicated to schools, if that is the more appropriate response to the face validity findings.
78. However, substantive doubts persist about the face validity of the test at the time of writing. Therefore, the achievement of this facet of the objective has not been demonstrated.

5.1.3 Content evidence of validity

79. Delegates to an RM teacher panel were shown the 3 – 5 and 4 – 6 tiers of the test. They were asked to state whether, in their opinion, the 2006 summative test addressed individual statements in the ICT programme of study.
80. The teachers considered that the majority of the PoS statements were covered by the test. There were 18 statements in total, 12 of them were considered covered for the lower tier, and 12 for the upper tier.
81. The statements that were not perceived to be covered by the test were not uniformly distributed across the PoS. Indeed, the statements that were perceived to be omitted included several that were intentionally not covered in the 2006 test.
82. Further, omitted statements tended to be in areas which required pupils to work (and especially to communicate) with others. This may confirm the view that it will always be difficult to cover such aspects of the curriculum in a 'closed-book' type test.
83. However, it has been countered that the test does simulate working and communicating with others – it is just that such work and communication tends to be remote (e.g. by receiving emails, carrying out instructions in them and then sending replies).
84. It should also be noted that existing NC tests sample from the PoS, rather than purporting to cover the whole curriculum every year.

© 2006 QCA

5.1.3.1 Evaluation of content validity

85. The findings reported here, whilst based on a small sample of teachers, reinforce findings from previous years. As such it can be said that content evidence of validity has been demonstrated.

5.1.4 Reliability

5.1.4.1 Definitions and methods

86. Reliability is a fundamental property of effective measurement. The reliability of scores derived from a test can be conceptualised in two basic ways. Firstly, a reliability co-efficient is a description of the amount of variance in scores that can be accounted for by the ability that the test purports to assess. This concept can be expressed through an analogy as follows: reliability is the extent to which test scores represent 'signal' rather than 'noise'.
87. The second way to describe reliability is as replicability in measurement. A reliable test is one that tends to provide a consistent measurement on repeat administration – assuming other conditions remain consistent between administration (e.g. test takers' motivation, no learning of content or familiarisation effect between administrations, etc.).
88. Experiments to establish the reliability of test data are of several types. Essentially, reliability experiments demonstrate the extent of consistency in results from parallel forms of the test in question.
89. Many traditional test designs use a form of analysis known as 'internal consistency reliability analysis'. Such analyses examine many different 'splits' of the data file produced from a test administration and seek to establish the average consistency of the two halves of test data.
90. 'Internal consistency' or 'split-half' reliability analyses have the considerable practical advantage for test developers that they can be derived from a single administration of a test.
91. However, there are several disadvantages for such types of analysis. Firstly, internal consistency analyses depend upon test data which has a fairly constrained type of structure. This – by extension – tends to require tests to be quite conservatively designed (for instance, based on many discrete-point items). The KS3 ICT test data are not structured so as to easily employ an internal consistency approach to reliability analysis.
92. Indeed, the use of internal consistency reliability indices for the KS3 ICT test has required researchers to create an 'approximation' to conventional analyses, by

© 2006 QCA

constructing 'task scores' specifically for the purpose. This limits the extent to which the internal reliability indices derived for the KS3 ICT test can be compared with existing reliability indices.

93. Despite this limitation, it is still intended to compare KS3 ICT internal consistency and that of other NC tests. This is done to try to understand the emerging picture of the ICT test's reliability, rather than to take a definitive view.
94. The second major disadvantage of such internal consistency analyses is that output reliability indices can be difficult for test users to directly interpret with respect to a national curriculum test. For instance, if a test that has an internal consistency index of 0.9, a user would not be able to straightforwardly establish its propensity to classify pupils consistently into the same level on repeat administration.
95. Thus, the second method for reliability analysis is known as the 'test-retest' method. In this method a sample of pupils are asked to sit a form of a test on two occasions.
96. The test-retest method has the advantage of being the most suitable for the type of data produced by an innovative test like the KS3 ICT test. Also, test-retest is most commonly associated with indices that describe the test's propensity to classify pupils into the same level on repeat administration. Such indices have the potential to describe the test's reliability very directly for test users.
97. Test-retest has some disadvantages, however. These are mostly practical in nature. For example, it can be difficult for test-retest studies to recruit sufficient numbers of schools and pupils to participate (because sitting two test administrations amounts to a significant commitment on the part of schools and pupils for no perceptible benefit). Also, there is the issue of the extent to which pupils' performance between the two administrations remains constant; for example, motivation may drop off during the second administration, or – conversely – pupils may become more familiar with the test environment during the first test administration, or have been taught more ICT and thus score more highly the second time around.
98. Despite these practical concerns about test-retest, the project focused on test-retest and classification consistency as the key issues in reliability analysis, but also ran internal consistency measures on the main administration of the test – this was taken as a 'back up' analysis, not a definitive statement of the test's reliability.

© 2006 QCA

5.1.4.2 Results

5.1.4.2.1 Internal consistency indices

99. Internal consistency reliability analysis was conducted on the full data set for the 2006 summative test. The analysis derived Cronbach's alpha[1] indices for several entities. These were the scores that allowed the analysts (in the lower tier data file) to calculate whether a pupil should be awarded level 3 (the L345 score), level 4 (the L45 score) and level 5 (the L5 score). Similar scores existed for the upper tier: L456 (to award level 4), L56 (to award level 5) and L6 (to award level 6).

100. Cronbach's alpha was calculated for form A and its clone sibling form B – see paras 391ff.

101. The tables (below), which are copied directly from a project draft validity report, show reliability values for the 2006 summative test, and juxtapose them with similar values derived from 2005 and 2006 pre-test data sets.

	2006 summative test		**2005 summative test**	**2006 pre-test**
Score level	Form A	Form B	Form A	Form A
345	0.779	0.786	0.783	0.659
45	0.742	0.752	0.700	0.600
5	0.629	0.646	0.498	0.497

Table 2: Cronbach's alpha for the lower tier

	2006 summative test		**2005 summative test**	**2006 pre-test**
Score level	Form A	Form B	Form A	Form A
456	0.690	0.702	0.798	0.592
56	0.592	0.607	0.649	0.578
6	0.405	0.365	0.116	0.428

Table 3: Cronbach's alpha for the upper tier

102. There are two types of comparisons that one might make in evaluating these reliability co-efficients. Firstly, it is useful to see if the 2006 summative test reliability co-efficients have increased or decreased – compared with their antecedents. The second form of analysis would be to compare these reliability co-efficients with those from other tests, or recommended in authoritative documents.

[1] Lee Cronbach was the psychometrician who developed this approach to reliability analysis.

© 2006 QCA

103. A useful way to see if reliability values have increased or decreased is to plot them on line diagrams. Figure 1, below, compares reliability coefficients for the lower tier of the L3 – 6 test.

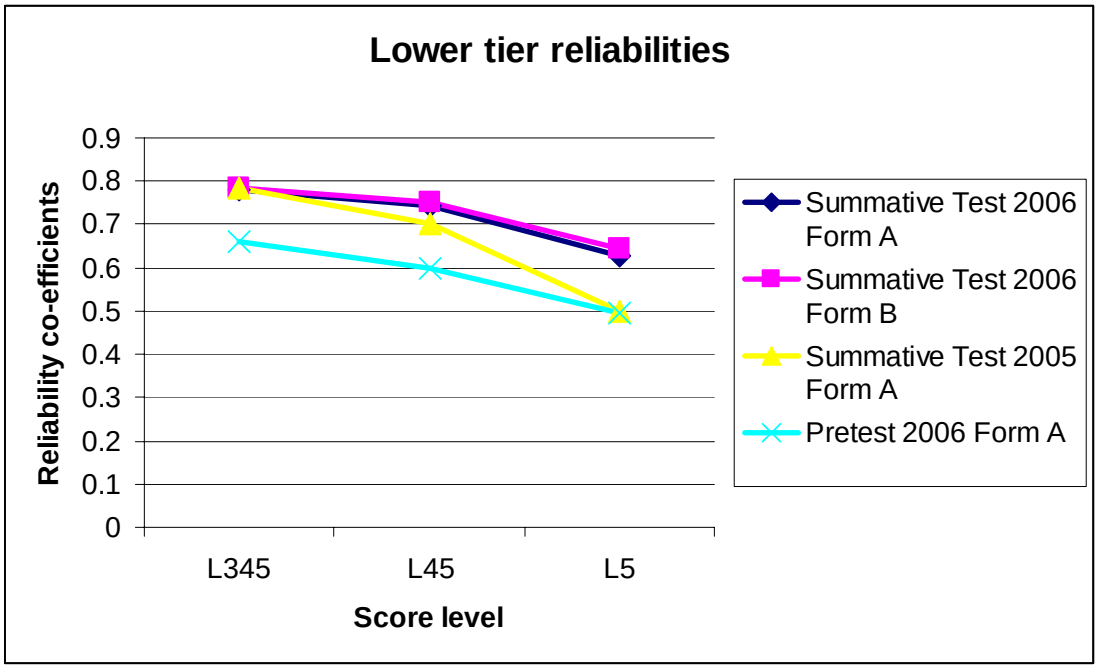

Figure 1: Comparison of reliability co-efficients for lower tier of test

104. This plot shows that the reliability co-efficients for the 2006 summative test are the highest of the four values reported (with the exception that the 2005 summative test L345 score returned a slightly higher reliability co-efficient than its 2006 counterpart).

105. However, despite the fact that the 2006 summative test co-efficients were the highest, the difference between 2006 summative and other values was quite small.

106. Upper tier values for Cronbach's alpha are shown in the figure below:

© 2006 QCA

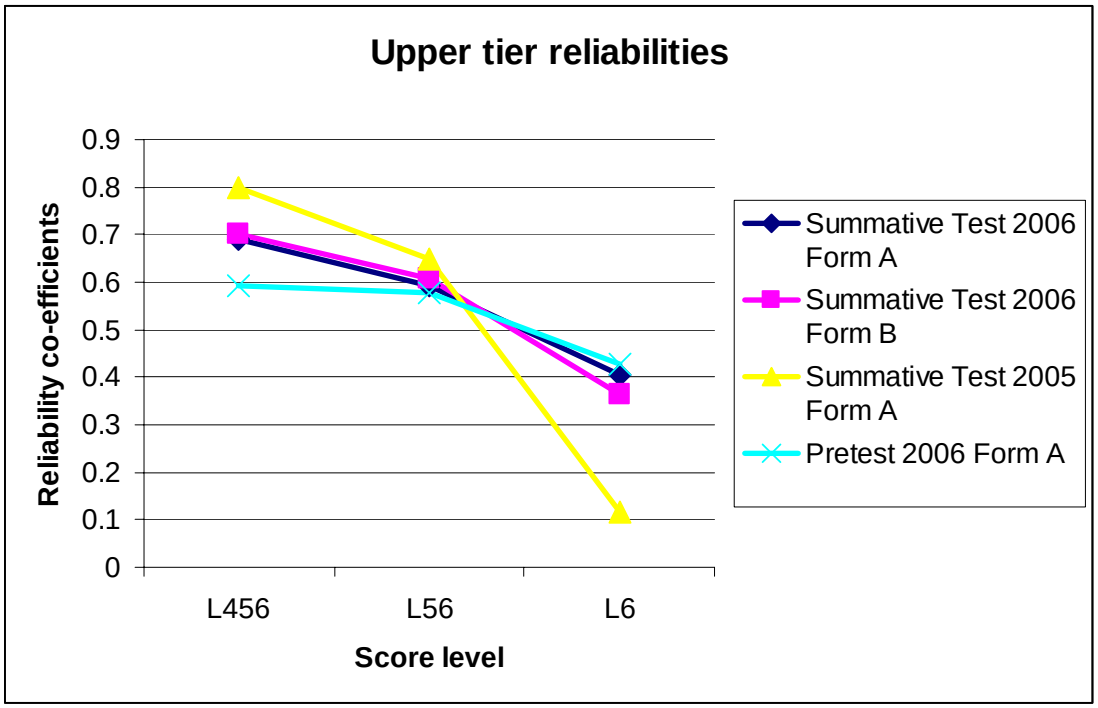

Figure 2: Comparison of reliability co-efficients for upper tier of test

107. Here the lines cross over – the 2005 summative test returned a higher value for reliability for the L456 and L56 scores than the 2006 summative administration. However, both administrations and both forms of the 2006 test displayed considerably higher reliability co-efficients than the 2005 test for the L6 score.

108. Thus, the internal consistency reliability findings demonstrate that some improvement has been made in reliability compared to the 2005 test. However, this is quite small improvement, and is not universal across all the levels that the tests address.

109. The next issue is to compare the reliabilities derived for the KS3 ICT test with those recommended in the literature, or derived from other tests.

110. Although there is a large body of literature emphasising the importance of reliability for high-stakes tests, there are surprisingly few statements of what acceptable values for reliability co-efficients for high-stakes testing might be. The highly-regarded American publication *Educational Measurement* states that a reliability co-efficient of 0.7 or greater is necessary for a high-stakes test. Other sources have suggested that high-stakes tests should reliability co-efficients of at least 0.8.

111. The other sources of evidence for the reliability values that might be expected for this test would be instruments that had similar functions. Table 4 shows

© 2006 QCA

Cronbach's alpha values for the three national curriculum tests that were administered on a statutory basis at key stage 3 in 2005.

Subject	Tier, etc.	Paper/Form	Cronbach's alpha values
English		Reading paper	0.81
		Writing paper	N/A
		Shakespeare paper	N/A
Science	Tier 3-6	Paper 1	0.92
		Paper 2	0.93
	Tier 5-7	Paper 1	0.88
		Paper 2	0.90
Mathematics	Tier 3-5	Paper 1	0.902
		Paper 2	0.914
	Tier 4-6	Paper 1	0.881
		Paper 2	0.868
	Tier 5-7	Paper 1	0.896
		Paper 2	0.888
	Tier 6-8	Paper 1	0.888
		Paper 2	0.886
	Mental test	Test A	0.896
		Test B	0.879
		Test C	0.850

Table 4: Cronbach's alpha reliability co-efficients for 2005 KS3 tests

112. The table shows that alpha values for maths and science tend towards 0.9 – the lowest being 0.85, and the highest 0.93. English is somewhat lower; the reading paper being slightly above 0.8. Reliability co-efficients were not calculated for the other English papers, which did not contain items with 'right or wrong' answers.

5.1.4.2.2 Classification consistency findings

113. A test-retest exercise collected usable data from 407 pupils in six schools.

114. The levels that pupils in the study achieved on first and second sitting of the test are recorded in the following table:

		Retest level					Total
		N	3	4	5	6	
Test level	N	2	0	4	0	0	6
	3	2	21	7	0	0	30
	4	6	10	74	58	3	151
	5	1	1	32	137	16	187
	6	0	0	1	11	21	33
Total		11	32	118	206	40	407

Table 5: Cross-tabulation of the levels achieved in the summative test and retest

© 2006 QCA

115. The table shows that 255 pupils (63 per cent) obtained the same level in the summative test and the retest. Of the 152 pupils who did not achieve the same level on both sittings, 64 did worse in the retest and 88 did better when retested.

5.1.4.3 Discussion of reliability results

116. Thus, the following may be said in the light of all reliability findings available to date.

117. Reliability is an intrinsic property of effective measurement; in fact it is a *sina qua non. Un*reliable measurement would be an oxymoron.

118. There is no value for co-efficient alpha that can be definitively taken to guarantee good measurement. However, it should also be noted that the existing national curriculum tests return higher alpha co-efficients than those derived from this administration of the KS3 ICT test.

119. There are three senses in which it might be open to supporters of the KS3 ICT test to argue that lower internal consistency co-efficients than are achieved for other tests with similar purposes are acceptable. These three senses are:

- The 2006 test was a pilot. The other tests are long established, and their reliability co-efficients should be interpreted in that light. There is some (slightly mixed) evidence that reliability co-efficients for the KS3 ICT test are 'getting better' year on year.
- The structure of the data produced by the KS3 ICT test is not intrinsically suited to internal consistency analysis. Since the test is made up of multi-faceted tasks, requiring several related sets of abilities, it is not surprising that the data set produced by pupils interacting with it is less internally consistent than data sets produced in conventional tests.
- There are precedents for tests containing papers or sections which by design have relatively low reliability, but which are held to maintain the test's validity in other ways. For example, some reputable foreign language examination batteries now contain writing and speaking papers to supplement multiple-choice-based elements, even though the writing and speaking papers tend to reduce the overall reliability of the test. Also, Table 4 shows that the writing and Shakespeare papers are present in the KS3 English test – even though they do not report a reliability co-efficient at all.

120. A test-retest study showed that 63 per cent of pupils were awarded the same level on retest. The meaning of this initial statistic is not yet clear; for example, the impact of factors such as diminishing motivation (tending to decrease retest performance) or increased familiarity with the test environment (tending to increase performance) are not known.

121. In further work, the project intends to calculate a statistic (Cohen's kappa) which can represent the extent to which pupils are consistently classified into levels. Kappa is a more robust measure than simple percent agreement calculation since it takes into account the agreement occurring by chance.

© 2006 QCA

122. Following this analysis, project researchers intend to simulate kappa statistics from reliability information held on the existing national curriculum tests.

123. Additionally, it will be very important to make sure that the 'experiment' that is set up to investigate classification consistency is carefully designed. If the experiment is not well designed, the true meaning of findings might be obscured.

124. The implementation of such a carefully-designed experiment should provide a set of statistics that will allow the tendency of the KS3 ICT test to classify pupils consistently to be evaluated in a principled manner.

5.1.4.4 Evaluation of reliability

125. It is clear that a substantial amount of high-quality work has already been carried out, and that more work is planned to investigate the reliability of test scores, and awarded levels.

126. However, at the time of writing it is not possible to make a completely watertight argument that the KS3 ICT test provides reliable measurement. As such, the achievement of this part of the objective has not been demonstrated.

5.1.5 Fairness for all pupils

5.1.5.1 Definitions

127. It is important that persons interpreting test scores can be reassured that the variance in scores that they interpret occurs as a result of pupils' abilities in the subject being assessed. Deviations from this desirable situation might occur if test data contain variance that results from random error (this is domain of reliability – see para 86 above), or if the pupils' interactions with the test interface were systematically affected by some source of difficulty that was not intended by the test developers and which was not part of the stated construct to be assessed (see para 376).

128. In contrast, the issue of fairness for all pupils is relevant to any situation in which an identifiable group of pupils could be seen to be unfairly disadvantaged in their experience of the test. The current QCA regulatory common criteria put an obligation on test developers to minimise bias. This evaluation prefers the positive construction that the test should be fair for all pupils.

129. The independent panel that reviewed A level procedures in 2002 gave the following useful definition of fairness for all pupils:

> 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

© 2006 QCA

> 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.

5.1.5.2 Results

130. Analysis was undertaken to compare the performance of boys and girls in the test. Opportunity counts from the lower tier of the test showed that girls scored more highly on the test than boys did. Similarly, girls scored more highly than boys in the upper tier.

131. The higher scoring of females was consistent with the evidence from Initial Level Assessments; the teachers' view before the test was that girls tended to be of higher ability than boys.

132. Further, previous evidence (such as the 2005 evaluation) did not throw up significant concerns about the fairness of the test for either gender.

133. Thus, the judgement of this evaluation is that current evidence suggests that this test is fair for both genders.

134. The performance of pupils who were entitled to Free School Meals (FSM) was analysed on both tiers. In the lower tier FSM-entitled pupils scored more lowly than those who paid for their own meals. Similarly, in the upper tier pupils who were not entitled to FSM scored more highly than those who were.

135. As in the case of fairness for genders analysis, the lower scoring of pupils who were entitled to FSM was consistent with teachers' prior judgement of their abilities in ILAs. Also, there is considerable evidence from various levels of national testing and public examinations over several years that pupils who are entitled to FSM do tend to score more lowly than those who are not.

136. Thus, this evaluation finds that the best current interpretation is that the test was fair for pupils who were entitled to free school meals.

137. Initial analyses on the performance of pupils with statements of Special Educational Needs found that such pupils had scored fewer opportunities than pupils without SEN statements.

138. The low scoring of pupils with SEN statements was consistent with the lower ILAs that such pupils received. However, further analysis will be conducted to seek reassurance that the test is fair for pupils with SEN.

© 2006 QCA

139. The situation for pupils who speak English as an Additional Language (EAL) is similar to that for pupils with SEN. Pupils who speak EAL scored fewer opportunities on the test than those who speak English at home. This was consistent with other indicators of their ICT ability (i.e. their ILAs).

140. However, the project intends to conduct further analysis to confirm that pupils with EAL are not being especially disadvantaged by some aspect of this test.

5.1.5.3 Evaluation of fairness for all pupils

141. The best current interpretation of analysis is that the test is fair for both genders, and for pupils entitled to free school meals. However, there is still work to do to fully establish the fairness of the test for pupils who have SEN or those who speak EAL. It is understood that this work is to be completed imminently.

142. Thus, there is partial, but not complete, evidence that this test is fair for all pupils. This state of affairs is consistent with a finding that the fairness of the test for all pupils has not yet been demonstrated.

5.1.6 Concurrent evidence of validity

5.1.6.1 Definitions

143. Sceptical observers are entitled to demand strong evidence that the key stage 3 ICT test is measuring 'correctly'. One part of an answer to mollify any sceptics is to demonstrate that the test can measure reliably (see para 86 above).

144. Reliability, however, essentially involves the consistency of data produced by a test when compared with itself (either in a 'split-halves' or 'test-retest' experiment). It is legitimate that a test should be able to demonstrate that the results it produces are credible when compared against results produced by the same pupils using a credible external measure of the same or a similar construct.

145. A common method for deriving concurrent evidence of validity is to ask a sample of pupils to take the test being developed, and also to take an existing test of the construct which is addressed by the new test. This approach has been considered in the current context, but has not been pursued for practical reasons.

146. In the absence of a second test, concurrent evidence of ability has been derived by using teacher assessments of ICT ability, provided by teachers of pupils in the main study for the key stage 3 ICT test.

© 2006 QCA

5.1.6.2 Results

147. Teachers' assessments of specific pupils were matched with levels achieved in the test. They are presented below as a table and as a figure.

148. The table is as follows:

Level achieved in test	Teacher Assessment Level							Total
	2	3	4	5	6	7	8	
N	2	78	286	636	39	1	0	1042
3	7	235	948	137	2	0	0	1329
4	1	335	2429	3870	606	36	2	7279
5	0	126	1228	3857	1342	250	16	6819
6	0	2	16	689	580	226	22	1535
Total	10	776	4907	9189	2569	513	40	18004

Table 6: Cross-tabulation of pupils' TA with the level achieved in the test

149. The table data have been developed visually in Figure 3 (below). It shows the data from the table arrayed along *x* and *y* axes. The cones in the figure give a visual impression of the third dimension in the table – that is, the more cases in a particular cell in the table, the bigger the cone on the figure.

150. Two visual effects have been carried out to show the propensity of the test and TA to classify pupils into the same level. Firstly, each cone where TA and test level coincide has been given a thicker border. Secondly, the cone which represents the biggest number of pupils in each row (i.e. test level) has been shaded with a striped pattern.

© 2006 QCA

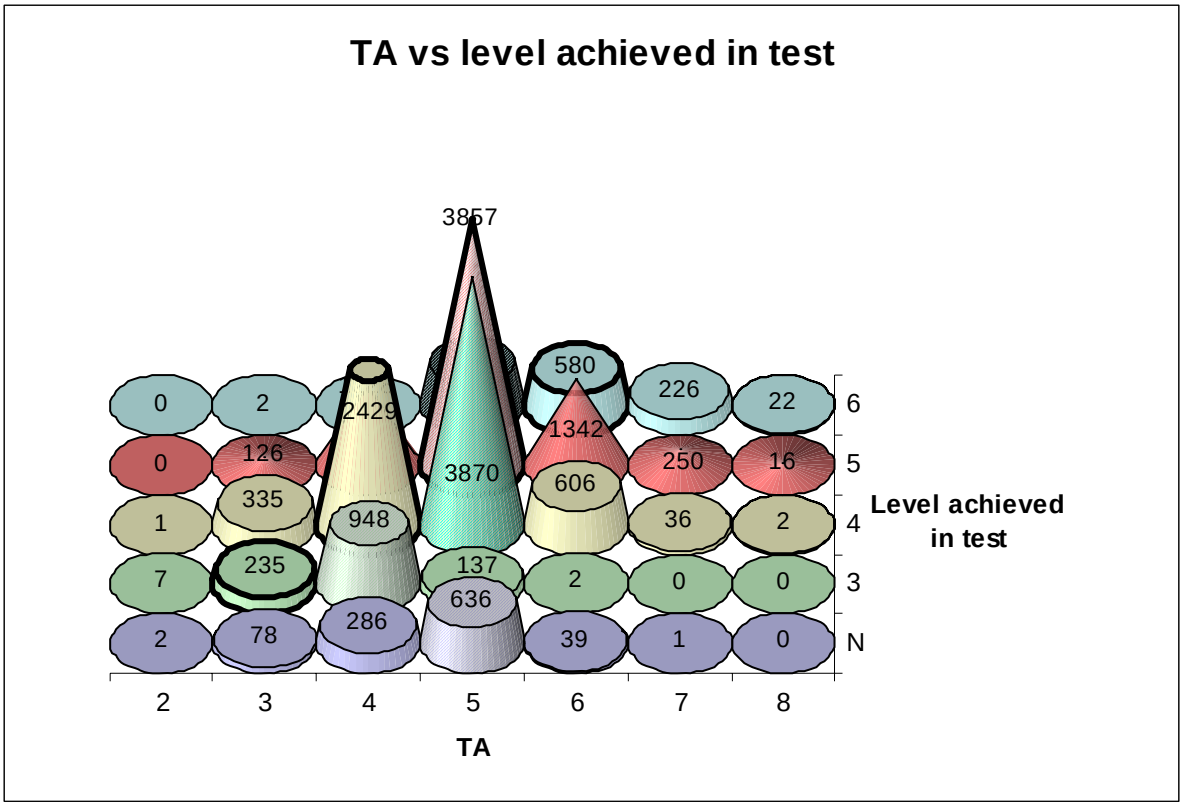

Figure 3: Comparison of numbers of pupils awarded TA and test levels

151. Taken jointly, the table and the figure show the following:

- There were 7101 pupils who were awarded the same level by test and TA from a total of 18,004 in the study. This is approximately 39 per cent of pupils in the study.
- This 'headline figure' might slightly underestimate the amount of agreement between the two methods of classification; for example, pupils who were entered into this test with levels 7 or 8 TA could only be awarded level 6 at the highest.
- Across the rows and columns, the cell that shows the 'matched' TA and test level – i.e. the one where pupils were awarded the same level – is the most populated cell (largest cone) more often than not.
- This is not always the case, however. For example, there are 1342 pupils with TA of 6, who achieved level 5 in the test; 606 who achieved a level 4 in the test; but only 580 pupils with TA level 6, who also achieved level 6 in the test.
- TA level 5 has the largest cones on the figure. There were 3857 pupils who were given a TA level 5, and who were also awarded level 5 on the test. However, slightly more pupils (3870) were given TA level 5, but got level 4 in the test.
- There were 1049 pupils (approximately 6 per cent of the matched sample) whose test result was at least two levels below their TA result (excluding pupils receiving level 'N' in the test and pupils with level 6 in the test, but level 8 in TA).

© 2006 QCA

5.1.6.3 Evaluation of concurrent validity

152. Concurrent evidence of validity has not yet been fully demonstrated; there is some evidence from the external measure to support the validity of the test's classifications into levels, but there is also substantive evidence that supports the contrary conclusion. It is of particular concern that approximately six per cent of the small matched sample's test score was two or more levels below their TA.

153. It is understood that more research is being carried out to evaluate concurrent validity. It is suggested that such research needs to produce a statistic to illustrate the consistency of classification (probably Cohen's kappa, as in the reliability case – see para 121 above).

154. When such a statistic has been derived, it will be possible to compare the instances of same-level classifications between the ICT test, and the more established NC tests. However, it may in fact be more problematic to evaluate the relationship between TA and test in the existing NC tests, in that they may not be genuinely independent – teachers generally know test results for pupils before they submit their TA. As such, it may even be the case that the seeking of concurrent evidence of validity is ultimately fruitless in this context – it is perhaps more helpful to look for construct evidence to support the validity argument[2].

5.1.7 Construct evidence of validity

5.1.7.1 Definitions and methods

155. The results of concurrent validity analysis were somewhat inconclusive. Thus, it is also useful to do wider analysis into the construct assessed by the KS3 ICT test.

156. Construct validity can have many definitions. For some writers, construct validity is equivalent to the overall definition of validity. However, the project specification defined construct validity as follows:

> Construct validity is a judgement of how well the assessment, or test in this case, calls upon the knowledge, skills and understandings of the construct, or constructs, that are the targets for the assessment. It requires a clear definition of the domain being assessed and evidence that, during their test 'performance', the intended skills and knowledge are demonstrated by the candidates.

[2] For comment on the validity (and alleged systematic leniency of TA in ICT), see para 197 in the level-setting section of this report.

© 2006 QCA

157. Further construct validation will be done on data derived from 2006 national curriculum tests (for example by comparing how pupils performed on some aspects of the key stage 3 ICT test with their performance in aspects of other NC tests where, it is hypothesised, they should have demonstrated similar performance).

158. One piece of work that has been conducted in 2006 could be considered to be construct validation. This has been the sources of difficulty project, which was carried out by researchers from the University of Leeds.

159. The premise behind this project was that in conventional testing the sources of difficulty that might affect test takers are well known. For example, teachers and pupils will know what makes an essay prompt or a multiple-choice question difficult (e.g. pupils might have learned how to spot distracters). In contrast, it was posited that the sources of difficulty that pupils could experience in a sophisticated interactive e-test would not be well known.

160. It was further posited that some of the sources of difficulty in the test might not be intended by the test developers. If that were so, it is possible that such unintended sources of difficulty might be an illegitimate source of variance in the test data, and – as such – a source of invalidity.

161. The Leeds research used qualitative methodology – by observing test sessions and interviewing teachers and pupils. The intention was to propose a map of the 'terrain' of sources of difficulty, and to suggest some potential sources of difficulty that might be in the test. Being exploratory, this research could not definitively account for the existence or absence of sources of difficulty based on a large sample of data.

5.1.7.2 Findings

162. The Leeds team posited a set of sources of difficulty that might affect pupils during the test. This list of sources of difficulty was organised into a taxonomy[3]. Principal dimensions of that taxonomy included whether the source of difficulty was related to the tasks in the test or to pupils' preparation for taking the test.

163. At the time of writing Leeds have passed on interim findings from the research to the main test development team in the form of a set of suggestions for how to mitigate unintended sources of difficulty in the test.

164. The following list is a selection of Leeds' suggestions:

[3] That is, a list grouping sources of difficulty and showing relationships – designed to suggest the main features of the terrain of sources of difficulty.

© 2006 QCA

About the tasks

- Find a way to make as explicit as possible what kind of activity is sought, and what is not wanted.
- Start each task with an easy part to get the pupils involved.
- If something is difficult, give credit for it. If credit cannot be given for it, make it easy.

About the test writing process

- Use story boards or similar ('cognitive walk through') to generate a narrative of what might be done in practice, to try to anticipate where pupils might get bogged down, and then build in support or preventative measures to improve likelihood of sustained engagement.
- Allow space within test development for the digitised question to be looked at by the item writer.

About the software

- Put plain language in the software menus, and increase the labels and descriptions attached to them, e.g. do not retain the formal approach of generic software such as the exotic language of databases, but write it in common sense terms, and/or offer direct pop up help for interpretation.
- Give more help to get out of 'holes', e.g. pop-ups that explain what should be done to escape, and have recovery/restore buttons for catastrophic mistakes.

About giving advice to schools about preparation and familiarisation

- Advise schools about best preparation – e.g. to include focus on the use of the toolbar, and the maximising of screens to aid organisation and visual clarity.
- Include within familiarisation guidance something that would enable development of test awareness related to good performance.

165. The test development project has interacted with these interim findings (including questioning some of them, and asking for further clarification). The Leeds researchers will present a final report to QCA at the end of September 2006, and it is understood that the test development project will adopt as many of Leeds' suggestions as is practicable, and/or consistent with other constraints (e.g. the need to deliver live tests to tight timetables, and the desire of the project to move into a phase of 'stabilisation').

5.1.7.3 Evaluation of construct validity

166. By design, much of the analysis into construct evidence of validity remains to be carried out. As such this facet of the validity objective has not yet been demonstrated.

© 2006 QCA

5.1.8 Level setting methods and outcomes

5.1.8.1 Overview of principles behind level setting

167. A detailed description of level-setting methods is given in Appendix B (see paras 398ff). This sub-section summarises some of the main points of principle that affect this aspect of the KS3 ICT test.

168. The levels that are awarded for national curriculum tests need to have two essential attributes: firstly, the distribution of levels needs to be set to provide a credible description of pupils' abilities in the assessed subject. Then, once credible levels have been set, the original standard needs to be maintained; essentially, so that comparable awards in different years can be demonstrated to represent comparable demands on test-taking pupils.

169. The two processes of standards setting and maintaining are quite different. Some of the key attributes of each process in the context of the KS3 ICT test are set out below.

170. The initial establishment of a standard is normally conducted using expert judgement. In such procedures, judges are often shown test scripts (which may be ranked by score, for example), and are asked to judge where thresholds should be placed in order to establish the standard.

171. Standards maintaining is a quite different exercise. In standards maintaining in national curriculum testing, it is normal for statistical techniques to also be employed to make sure that it is neither easier or harder for pupils to be awarded a particular level in any given year.

172. Whilst much attention in standards maintaining is focused on the comparability of standards over time, it is also important that levels awarded in different manners (e.g. levels that are common to different tiers; or that are awarded from different test forms) have comparable demand.

173. A point made elsewhere is repeated: the 2006 pilot of the key stage 3 ICT test did not amount to a statutory administration. The implication in this particular context is that the 2006 levels did not in any sense set a standard for future years to adhere to.

174. Further, when an initial standard for the KS3 ICT test is set, it will not be required to equate to any existing standard in ICT. This is despite the prior existence of distributions of levels awarded for ICT by teacher assessment.

175. Whilst there is no formal requirement on the test to equate to the levels awarded by TA, it is possible that some stakeholders might choose to interpret any discrepancies that might exist between TA and test-awarded levels as

© 2006 QCA

evidence of the invalidity of the new set of levels. The legitimacy of any such interpretations is discussed at para 198 below.

176. Further, the existence of NC level descriptions cannot be said to provide a prior standard to which a new test could be required to equate to. This is because level descriptions are in fact only 'best fit' descriptions of what a typical pupil at a given level would be likely to do. They are not in fact definite criteria for the award of levels.

177. The KS3 ICT test project has developed a novel method to award levels. This method is referred to as the 'sufficient evidence model'. The sufficient evidence model is described in more detail at paras 400ff.

178. In the 2005 pilot, there had been a discrepancy between the distribution of levels awarded by the 2005 pilot test and by TA. At the request of the QCA Executive, an independent panel of assessment experts was set up to review the levels awarded in 2005, and to make recommendations for how the level-setting procedures might be improved for future years. Since the level-setting procedures adopted in 2006 were quite similar to those of the previous year, key findings from the independent panel will be summarised in the next sub-section of this report.

179. As mentioned in Appendix A to this report (see para 371), the 2006 pilot benefited from the participation of QCA and NAA teams involved in level setting for other NC tests. As such, the level-setting meeting was overseen by NAA senior officers: the Director of Quality Assurance and the Head of Strategy and Policy.

180. Staff from QCA's Regulation and Standards Division also observed the main level-setting meeting. In addition, staff from this team attended the teacher panel that provided a set of draft level thresholds. The Regulations and Standards staff have written an informal report documenting their observations on these two aspects of the level-setting process.

5.1.8.2 Opinions on procedures

5.1.8.2.1 Independent panel findings on level-setting process and procedures

181. The independent panel that reported on the 2005 level awards concluded that awarding procedures were sound in principle. However, it also found that there were 'flaws' in the 2005 test and its delivery. But, it further stated that none of these flaws was so serious that it could not be remedied if actively addressed before 2008. The panel also made 18 recommendations to improve the test.

© 2006 QCA

182. Perhaps the key recommendation of the independent panel was the following:

> There should be no role for the test contractor or project team in setting educational standards – this should be a matter of professional judgement moderated by the QCA.

183. However, the NAA project team contested the independent panel's recommendation, and responded as follows:

> We disagree. The project is still in pilot and standards need to be set through us working in close collaboration with teachers. Once those standards have been established then our role will be to maintain them.

5.1.8.2.2 Issues from a Regulation & Standards team report

184. NC assessment regulation staff have questioned the focus of the teacher panel. The panel's work has been described as similar to 'a script scrutiny' exercise. However, the Regulation team's informal report questions whether the teacher panel was in fact functioning as a teacher judgmental exercise, rather than a script scrutiny.

185. The following table has been developed by regulation staff – in order to define and to mutually contrast script scrutinies and teacher judgemental exercises (TJEs):

© 2006 QCA

Process	Script scrutiny	Teacher judgemental exercise
Participants	Senior markers	Teachers
Chair	Marking Programme Leader – MPL (appointed to position by Test Operation agency, independent from Test Development Agency – TDA or NAA)	Representative from TDA (though not actively involved in the decision making)
Item under consideration in meeting	Scripts – i.e. a test paper with pupils' written responses.	Test questions – without any pupil responses
Timing	After the live administration of the test	Six months before the live test administration, using questions from final versions of test papers.
Selection of Panel	Part of responsibility as a member of the senior marking personnel	Invited by TDA from schools participating in pre-tests
Experience of test standard	Developing materials for marker training (up to six months involvement), and then marking own allocation of scripts	Shown test for the first time on the day
Presentation of results	By MPL directly to final level setting meeting	To draft level setting meeting by TDA personnel as part of the draft level setting report (which also includes equating data)
Impact of results	Potentially determines decisions made at final level setting meetings	One strand of evidence used to set script scrutiny ranges[4] and draft level thresholds

Table 7: Comparison of script scrutiny and teacher judgemental exercise

186. Thus, a strength of the script scrutiny procedure was that it was generally conducted with participants who had an intimate knowledge of the standard to be set and properties of performance at relevant thresholds. It was suggested that building up (perhaps over several years) a group of such experts to contribute to the KS3 ICT test standards setting process would enhance the robustness of standards set on the new test.

187. Further, the use of opportunity reports in the KS3 ICT teacher exercise is unusual, in that opportunities are neither pupils' responses to test questions, nor the questions themselves, as is the case in the two established methods described above.

[4] The mark range used to identify scripts as part of the level-setting process.

© 2006 QCA

188. The contrast of script scrutiny and teacher judgemental exercises is particularly important at the present time, as there is currently a review of TJEs, with a view to discontinuing them, due to their lack of robustness.

189. In contrast to the concerns expressed by regulation staff, the NAA officer who attributed draft cut scores to scripts reported that he found that process very intuitive and that he believed that it had contributed to a robust level-setting process.

190. Additionally, RM pointed out that teachers in the panel had said that they had found the process credible. Also, the three separate ratings of cut points only diverged by a relatively small amount. RM took these two facts as confirmation that the process produced consistent results[5].

191. The use of opportunities to set levels has also been questioned. The originators of the sufficient evidence model (see para 400 below) maintain that pupils achieving more opportunities cannot be said to have more ability in ICT than those triggering fewer opportunities – the presence of more opportunities is said to be merely evidence that allows an interpreter to have more confidence that a pupil is at a particular level.

192. Regulation and Standards commentators believe that this assertion is inconsistent with the summing of opportunities to provide cut scores.

193. The project has attempted to defend the notion of summing opportunities to provide cut point thresholds. They believe that the notion of ascending confidence in a level award with ascending numbers of opportunities is entirely plausible.

5.1.8.3 2006 level distribution

194. The percentages of pupils awarded specific NC levels are shown in the table below:

Test level	Percentages of 2006 pupil cohort
N	6.3%
3	7.0%
4	42.4%
5	37.1%
6	7.2%

Table 8: Final distribution of pupils into levels in the 2006 summative test.

[5] Caution should be exercised here, however. Since the cut scores are only relatively small numbers, it is perhaps to be expected that they will not diverge by much.

© 2006 QCA

195. The percentages of pupils allocated to specific levels by the 2005 and 2006 test and by TA in 2005[6] are shown in the figure below:

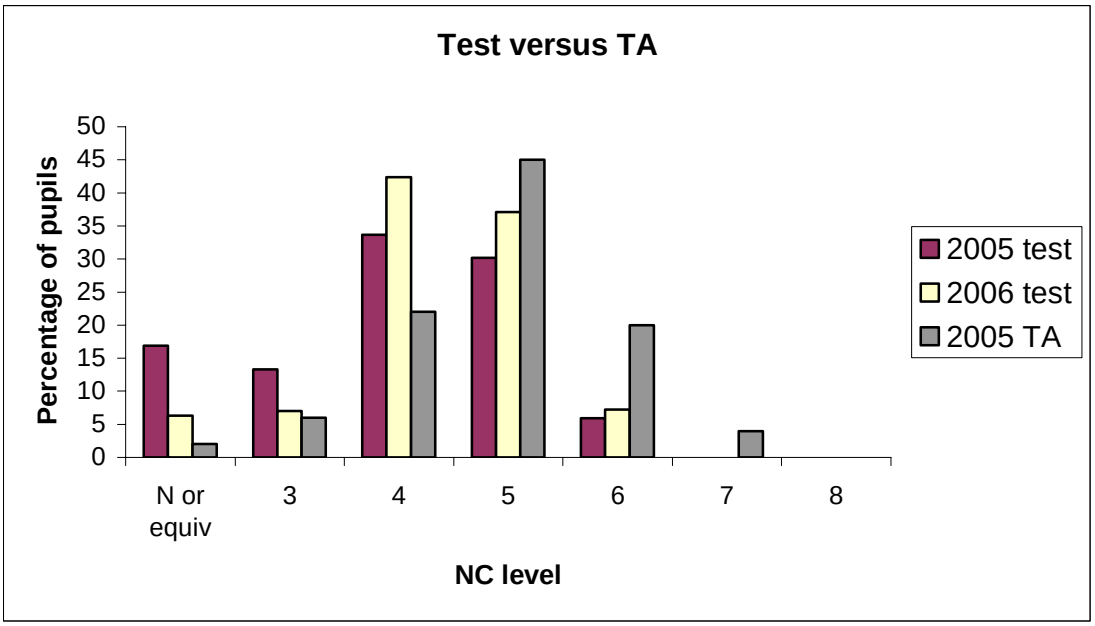

Figure 4: ICT levels awarded by test in 2005 and 2006, and by TA in 2005

196. Taken jointly, the figure and the table show:

- 44.3 per cent of pupils achieved level 5 or above in this test.
- The distribution of levels achieved in the 2006 test was somewhat higher than the distribution achieved in the 2005 test.
- The distribution of levels achieved in the 2006 test is still some way below the distribution of levels achieved in 2005 TA.

197. There has been some commentary (including excerpts from an Ofsted report) that – in previous years – teacher assessment in ICT was inappropriately lenient. As part of research for a briefing paper some ICT curriculum experts' views on the alleged leniency of TA in ICT were sought in August 2005. The experts' views included:

- We … don't believe that we have evidence to support [a] claim that TA assessed pupils at a level above their actual attainment.
- Previously … TA was up to a level too high. However, now things are much more accurate. Most teachers' assessments are pretty close to what pupils are actually doing.
- My estimate [a few years back] was that the whole picture was about one level out. The KS3 strategy certainly addressed this issue head on and I know that a lot of schools revised their practice in the light of the training they received.

[6] The 2005 TA figures are those submitted by teachers officially to the DfES; 2006 TA results are not yet available at the time of writing. The results in this figure are totals – they are not linked by pupils.

© 2006 QCA

198. It has also been observed – both by the briefing paper to the QCA Executive on the 2005 levels, and by the independent review of the 2005 levels, that the conditions do not currently exist to objectively evaluate whether divergent TA or test scores in ICT are the 'right ones'. Some form of controlled study would need to be established to investigate such a question.

5.1.8.4 Evaluation of level-setting procedures and results

199. The independent panel that reviewed the 2005 level awarding gave a general endorsement to level-setting procedures. In addition they made 19 recommendations for amendments to level-setting procedures, and to aspects of the test and its administration more generally.

200. The regulation staff who have observed level setting in 2006 have made some specific criticisms of awarding procedures. Similarly, concerns have been expressed as to whether the use of opportunities to support awarding can be defended.

201. The 2006 results were less divergent – both from TA results and from PSA targets – than the 2005 results. However, attainment as reported by the test remained substantially lower than in either of the other two cases.

202. It is understood that face validity research to be carried out in the next few weeks (see para 76 above) will ask teachers their views as to the credibility of level distributions.

203. Further, NAA is planning to run a technical seminar on KS3 ICT in autumn 2006. This meeting will invite a group of experts in relevant fields to examine in depth issues related to the measurement and awarding models used in the 2006 pilot. It is also understood that NAA intends to invite a consultant of international repute to further validate the models.

204. Whilst both the further face validity and technical investigations are to be supported, it remains the case that evidence that level-setting procedures are in line with best practice, and that the resultant distribution of levels is defensible has not yet been established.

5.1.9 Overall evaluation of validity

205. Validation of the 2006 pilot has taken many forms. Specifications for validity work were agreed early by RM and the NAA project team. Following those specifications, a wide range of research has been undertaken and reported openly in well-written reports.

© 2006 QCA

206. It is also known that substantial pieces of extra validation work are either currently underway, or planned for the next few months.

207. The evaluation of validity in the current report has been based on primary research into the 2006 pilot. It has required the test development project to actively provide evidence of validity in order for the test to be considered so. A summary of the 'micro-evaluation' of validity for each facet of the concept is shown in the table below:

© 2006 QCA

Facet of validity	Evaluation of facet	Outstanding issues	How covered
Face validity	Not demonstrated	Confirmation of decision to use a bespoke environment	Independent consultant appointed
		Opinion collection needing to be robust and transparent	Collaboration to improve opinion collection across the project
		Confirmation of questionnaire findings needed	Shorter, targeted teacher questionnaire to be administered in autumn 2006
Content evidence	Achieved	Investigation of whether ICT test actually covers more of curriculum than other NC tests.	
		Confirmation of whether test can adequately measure aspects of the curriculum related to 'communicating' and 'working with others'.	
Reliability	Not demonstrated	Derive statistic to show classification consistency for KS3 ICT and other NC tests	Research to be reported in 'additional validity report' in autumn 2006
Fairness for all pupils	Not demonstrated	Demonstrate fairness for pupils with EAL and SEN	Extra analyses to be reported in 'additional validity report' in autumn 2006
Concurrent evidence	Not demonstrated	Derive statistic to show classification consistency for KS3 ICT and other NC tests	Extra analyses to be reported in 'additional validity report' in autumn 2006
		Decide whether it is valid to use relationship between TA and NC test level for concurrent validity purposes	
Construct evidence	Not demonstrated	Integrate as many findings from sources of difficulty project as possible into working practices	Findings implemented in various ways – e.g. via amendment to task writer and reviewer guidance
		Analyse actual relationships between aspects of ICT test and other NC test performance	Research to be reported in 'additional validity report' in autumn 2006
Level setting	Not demonstrated	Confirm role of project and contractor in standards setting	Some disagreement over finding – 'statutory readiness process' and increased involvement of Regulator will ensure best practice is followed
		Confirm status of teacher panel as script scrutiny or teacher judgemental exercise	Issue to be addressed at autumn 2006 technical seminar
		Confirm validity of awarding levels on basis of opportunities as indicators of ascending confidence in a level award	Issue to be addressed at autumn 2006 technical seminar
		Get authoritative confirmation of sufficient evidence model.	NAA intends to appoint internationally reputable psychometrician to evaluate the model.

Table 9: Summary of micro-evaluations of facets of validity

208. The table makes plain that conclusive evidence has not yet been provided with respect to all but one of the facets of validity.

© 2006 QCA

209. Also, it is worth noting that there are validity issues not dealt with in this report that will need to be successfully resolved before the test can run successfully on a statutory basis. These include: task banking and the demonstration of an effective mechanism for maintaining standards over time[7].

210. It should also be noted that the project has plans in place to deal with many of the outstanding issues described in the table. Such work will help to understand the issues still to be resolved. However, the fact that such work is imminent does not in itself guarantee a successful result. For example, the reliability and concurrent validity analyses are by no means guaranteed to provide an unequivocally successful outcome.

211. The irresolution of substantial knotty questions is consistent with the ambition that informs this development; this project aims to roll out a highly sophisticated, interactive test to all schools in the country and to demonstrate its validity to stringent standards. It should come as no surprise that difficult questions still exist.

212. This situation with important outstanding work to be done to confirm the validity of the tests also needs to be understood within the temporal context. 2007 and 2008 test development work is already underway, and that pilot will need to provide clear evidence of the test's validity in order to facilitate a decision about statutory readiness for 2008.

213. Thus, taking into account the findings from the micro-evaluations of the facets of validity, and all the factors alluded to above, the finding of this evaluation is that validity has not yet been demonstrated.

[7] These issues are not evaluated in this report since the task banking and standards maintaining methods were not actively implemented during the 2006 pilot.

© 2006 QCA

5.2 Objective four

214. Objective four is:

> Provide all schools participating in the 2006 pilot with accurate formative reports from the practice test and an accurate summative report from the summative tests.

215. The Critical Success Factor associated with objective four is:

> This CSF is met if both pre-release testing and the use by pupils and teachers confirm that the formative and summative reports produced accurately reflect the activities undertaken and schools find the reports useful.

216. The other success factors are:

- Appropriate analysis, using an appropriate range of methods, and reporting on feedback from an appropriately sized sample of schools, demonstrates to QCA and DfES's approval, the usefulness and user-friendliness of the formative and summative reports, and the consistency of summative reports with the NC levels awarded by the test.
- The automated marking is generating statements for reports that accurately reflect what pupils have done.

5.2.1 Formative reports

217. The formative reports, and the conditions associated with their generation, are described in Appendix B (paras 405ff).

218. A research project into the formative reports was conducted by the writer of this evaluation report in the first half of 2006.

219. That research was based on three sources of evidence:

- Research literature into 'plain formative' and 'e-formative' assessment
- Documents produced by the KS3 ICT project
- Teachers' opinions, gathered in telephone interviews

220. The evidence from these three sources had limitations (for example, existing research into e-formative assessment tends to be less well-designed than that relating to plain formative. Also, despite the approach to a substantial number of teachers, only a relatively small number took part in interviews).

221. The weaknesses in two of the data sets were considered to be a factor that would limit the extent of claims that might be made from the data. However, the overall nature of the data was compatible with the aim and approach of this research; that is, an exploratory exercise to generate and exemplify important concepts relating to the formative reports.

222. Some key findings from each strand of this research are outlined below.

© 2006 QCA

5.2.1.1 Research literature into formative assessment

223. Formative assessment has been thoroughly researched in recent years and is now supported by a sound body of empirical evidence.

224. Some key findings on formative assessment have included:

- There was clear evidence that effective formative assessment could improve pupils' attainment.
- There was clear evidence of a poverty of practice in formative assessment – and hence a need to improve it.
- The ways in which teachers could improve formative assessment practice were clear:
  - Feedback to any pupil should be about the particular qualities of his or her work, with advice on what he or she can do to improve, and should avoid comparisons with other pupils.
  - For formative assessment to be productive, pupils should be trained in self-assessment so that they can understand the main purposes of their learning and thereby grasp what they need to do to achieve.
  - Opportunities for pupils to express their understanding should be designed into any piece of teaching, for this will initiate the interaction whereby formative assessment aids learning.
  - The dialogue between pupils and a teacher should be thoughtful, reflective, focused to evoke and explore understanding, and conducted so that all pupils have an opportunity to think and to express their ideas.
  - Tests and homework exercises can be an invaluable guide to learning, but the exercises must be clear and relevant to learning aims. The feedback on them should give each pupil guidance on how to improve, and each must be given opportunity and help to work at the improvement.

225. Also, researchers have suggested that written feedback should take the form of well-written, specific comments on work – but not marks or grades which (either explicitly or implicitly) compare the recipient's work (and by extension his or her self-worth) to that of his or her peers.

226. Some doubt has been cast on the practicality of implementing comment-only marking. However, the original researchers have countered that the doubts expressed about the original work do not amount to a credible rebuttal of their core findings, nor of the possibility of the findings' practical implementation.

227. In contrast to 'plain formative assessment', research into e-formative assessment is embryonic. It is often based on small data sets, much of it comes from higher education, rather than key stage 3, and at times, methodologies are not set out as clearly as they might be.

228. Despite these weaknesses, it was felt that there are some strands that are becoming clear in e-formative assessment research. These strands are summarised below:

- Formative e-assessment research has tended to be about assessment instruments used; many different test designs have been used to provide e-

© 2006 QCA

formative assessment, from multiple-choice tests to rich-media simulation-based assessments.

- Feedback delivered subsequent to e-assessments has taken many forms (e.g. right or wrong answers, marks or grades, rich media tutorials, etc.). Also, feedback has been delivered at many different junctures (e.g. within tasks or questions, after each question, at the end of each test sessions, etc.).
- Some researchers have claimed that feedback from e-formative assessments has benefits that are not available in pencil-and-paper assessment. For example, electronic media allow feedback to be presented in a variety of ways, and thus to be appropriate for learners of different cognitive styles.
- However, e-feedback could have disadvantages. For example, students have been observed to 'superficially' click through web site feedback, rather than engage in 'deep learning'.
- With some exceptions (e.g. e-portfolios used for formative purposes), formative e-assessment has tended to be associated with self- or peer-assessment. This is in contrast to the plain formative assessment research, which emphasises formative assessment as improved classroom interaction.

5.2.1.2 Project documents

229. Key benefit BS001 of the KS3 ICT programme is:

Immediate formative feedback from the practice test aids teaching and learning.

230. Original project documents envisaged the production of pupil-level reports and school-level formative reports. Whilst formative reporting systems are theoretically capable of producing school-level reports, currently only pupil-level reports are produced.

231. Between the 2005 and 2006 pilots, substantial work was carried out to improve the formative reports. This work included showing beta versions of the reports to teacher and national stakeholder groups, and then taking their suggestions into account to improve the reports.

232. Prior to releasing the 2006 practice test, RM conducted software tests to assure the accuracy of statements generated for formative reports. The testing attempted to establish that:

- formative statements accurately reflected what pupils had done in a practice test session.
- 'improvement statements' were both accurate – given what pupils had (or had not) done in the test – and logical (e.g. if a pupil had sent an email with an attachment, the report should not recommend that the pupil needs to work on his emailing).

233. RM concluded that this testing was successful, and communicated the formative reports' accuracy and logicality to NAA in its release recommendation for the practice test.

© 2006 QCA

234. Some questions in the 2006 pre-test questionnaire pertained to formative reports. However, the majority of teachers did not answer those questions.

235. Teachers' comments that were gathered from the pre-test questionnaire related to (amongst other things): the appropriateness of the language of the reports and the organisation of the reports into sections.

5.2.1.3 Telephone interviews

236. 246 schools were contacted with a view to carrying out a telephone interview into the formative reports. From the 246 contacted teachers, 27 telephone interviews were conducted.

237. This is a low response rate for a research exercise. It is, of course, very difficult to know why so few of the contacted teachers wished to take part in an interview.

238. Not all the 27 teachers had seen a formative report. 18 had seen either the 2005 or 2006 reports, and 14 had seen the 2006 reports specifically.

239. This meant that the data set provided by the telephone interviews was small – and this limits the interpretations that can be put on findings. However, although the data set was small, it did provide a sample of teachers' opinions. The data provided by the interview was mainly qualitative – each teacher answered up to 26 questions, the majority of which had a free-text response. This meant that each interview was a rich source of information.

240. There were some findings which were consistently expressed. These are listed below:

- Almost all the teachers believed that the formative reports should display a national curriculum level. However, when pressed to consider problems with so doing, their responses were more mixed; some maintained the view that levels could (and should) be reported, whereas other described problems with adding levels to reports.
- Many teachers commented on the language of the reports. This comment was mainly critical. The criticism had two strands: that the language was aimed at teachers and not pupils, and that the reading load was too demanding for year 9 pupils.
- Most teachers who responded thought that the reports were not currently targeted at the right audience; and that the correct target audience for the reports was children.
- However, there was a reasonably strong strand of opinion that argued that it was a teacher's job to mediate feedback to children, and therefore the audience for the current reports (teachers) was legitimate.
- Some responding teachers took the view that there should be formative reports designed for specific audiences: one for pupils, and two for teachers: one addressing an individual pupil, and a new type of report summarising the strengths and weaknesses of an entire group (either class or year group).

© 2006 QCA

241. There was a small amount of evidence from the phone interviews concerning the perceived accuracy of the reports. Firstly, teachers were asked whether the reports seemed to reflect what pupils had actually done in the practice test. Five teachers said 'yes' the reports did reflect what pupils had done, one said that the reports did not reflect what pupils had done in the test, but eight did not know whether or not the report reflected pupils' performance in the test. Secondly, teachers were asked whether the formative reports reflected what their pupils could do in normal class work. Here, three respondents said 'yes', four said 'no' and eight did not know.

5.2.1.3.1 Evaluation of formative reports

242. The following important facts are noted: the project responded to comment on the 2005 formative reports by implementing substantial improvements. These improvements were then checked via suitable forums – e.g. Teacher Review and National Stakeholder Groups.

243. The project ensured that formative reports were checked for accuracy before releasing the practice tests. This test appears to have been thorough, and returned a positive result.

244. When the formative reports were released, it would appear that they were little used by teachers. It is not possible – without a follow-up study – to ascertain why teachers did not use formative reports.

245. The wording of the CSF associated with objective four is:

> This CSF is met if both pre-release testing and the use by pupils and teachers confirm that the formative … reports produced accurately reflect the activities undertaken and schools find the reports useful.

246. Pre-release testing supports a contention that the reports were accurate. However, the best currently available evidence is that schools and pupils have not been using the formative reports. It further follows that they cannot find them useful. It is emphasised that this finding is despite the project's best endeavours to make them so.

247. As such, the achievement of this part of objective four has not been demonstrated.

248. The above finding concerns the formative reports as teachers have experienced them to date. However, the research into formative reports went further – making recommendations to improve formative reporting.

249. In repeating these recommendations, the following caveats are acknowledged:

© 2006 QCA

- The less-than-perfect data sample upon which the research was based means that these recommendations are a reasoned professional interpretation of limited evidence, rather than an incontrovertible finding from a large sample of data.
- There might be some feasibility issues – e.g. accommodating some of the recommendations with a test based on the 'sufficient evidence model'. Such feasibility issues would need to be resolved before commencing work.

250. The recommendations are:

- The development of a wider range of formative assessment materials – for example, a set or bank of tasks from which teachers could choose to suit their own formative assessment work.
- The production of three types of formative reports: a teacher's report on an individual pupil, a teacher's report on a group of pupils and a pupil's report. These would have specific design features, suitable to the audience and purpose.
- National curriculum levels should not be added to formative reports, despite the finding that many teachers would welcome this[8].

5.2.2 Summative reports

251. The summative reports, and the conditions associated with their generation, are described in Appendix B (paras 412ff).

252. Prior to releasing level 3 – 6 results to schools, RM prepared a release recommendation to NAA. In this recommendation, it was stated that due diligence and quality assurance had been undertaken to facilitate the results release.

253. The testing consisted of the following elements:

- Set-up of APSes in order to simulate five schools
- Generation of test data within the five dummy schools
- Upload of the test data to the CPS and subsequent validation
- Simulated moderation of the marks obtained
- Generation and examination of the test report on the CPS
- Release of results to the APS and subsequent validation
- Generation and examination of the test report on the APSes.

254. It is stated that the successful conduct of these tests amounts to a confirmation of the accuracy of the pupil reports.

255. It is understood that no evidence has yet been gathered of teachers' (and pupils') opinions as to the accuracy and usefulness of the summative reports.

[8] This is because of the strong message from experts in formative assessment that adding grades or marks to formative feedback diminishes the learning gains from such feedback.

© 2006 QCA

(The reports were sent to schools too close to the end of the summer term for an opinion-collection exercise to have been taken out yet.)

5.2.2.1 Evaluation of the summative reports

256. The evaluation for the summative reports is similar to that for the formative reports. There is good evidence that the reports accurately reflected what pupils did during the test. However, there is not evidence that teachers and pupils found the reports useful. In this case, this is a consequence of the very recent release of the summative reports.

257. As such, the evaluation of the summative reports is that this facet of the objective has not been demonstrated.

5.2.3 Overall evaluation of objective four

258. Further to the findings on formative and summative reports, the evaluation is that objective four has not been demonstrated.

© 2006 QCA

5.3 Objective six

259. Objective six is:

Ensure that schools register interest in the pilot, complete accreditation, move to Approved Test Centre status and with adequate preparation time, take part in the 2006 summative test window, and have a satisfactory experience throughout the process.

260. The Critical Success Factor associated with objective six is:

This CSF is met if:

- fewer than 50 maintained schools fail to register interest; fewer than 50 fail the technical accreditation; and fewer than 200 complete accreditation but do not take part in the 2006 summative test window (for the avoidance of doubt, this permits 300 or fewer maintained schools not participating in 2006); or
- if these targets are not achieved, a credible action plan is in place to ensure targets are achieved in the 2006/07 school year.

261. The other success factors are:

- 95% of schools that return feedback questionnaires report:
  - overall satisfactory experience;
  - positive feedback on contact with RM;
  - positive feedback on contact with Becta;
  - quality and helpfulness of support and training materials;
  - ease of administration of the test;
  - the running of the test to reveal no critical faults.
- RM and QCA's monitoring of schools' experiences is found to have been pro-active, sympathetic and effective. No serious widespread concerns arise for schools during the pilot without an appropriate branch of the project taking prompt and appropriate action to communicate with the schools concerned and, as far as possible, mitigate the reported problem.
- 95% of schools which have had contact with LA Strategy consultants express an interest in the test.
- Becta contact all schools which have not attempted, or have failed, the network audit.

262. Two observations about objective six are apt: firstly, this objective was developed in the early part of the 2005 – 06 pilot cycle. As such, the definition of the appropriate number of schools to participate in the 2006 pilot was carried out before the transfer of programme management responsibilities to NAA, and the instigation of the Wider-School Readiness project (see paras 373 and 313 respectively). Secondly, the number of non-participating schools permitted by the CSF was known to be 'a difficult target', but 'the bar was set high' deliberately, to reflect the importance that was attached to achieving full participation.

© 2006 QCA

5.3.1 Participation

5.3.1.1 Definition of schools eligible to take part

263. Defining which schools are eligible to take part in a pilot of a national curriculum test (and – by extension – will be required to take part in it, once it is statutory) is less straightforward than might, on initial consideration of the issue, seem the case.

264. The majority of schools for whom the test will be compulsory once statutory are straightforward; state-maintained secondary schools need to be considered when calculating participation figures.

265. However, small schools present a particular issue for participation purposes. Special schools and Pupil Referral Units may have only a few pupils in year 9 working at level 3 or above, and hence it may be that an onscreen test is not the most effective way to deliver assessment for such small groups. This observation (which remains a supposition) has considerable implications for participation purposes, however.

266. In addition to the issues above, there remain further complications when schools close, or merge, or when new schools are opened. In the case of new academies, the extent to which they will be *required* to enter all year 9 pupils for statutory NC tests is also a consideration, when attempting to decide how many schools should be counted when evaluating the participation in the 2006 pilot.

267. The table below shows the most accurate count of schools that is currently available. It follows that since the 'headline total' number of schools has changed several times in recent months, the total numbers of schools quoted in analyses below may vary.

© 2006 QCA

Date	Previous total schools	Current total schools	Reasons for change
28th April	3885	3880	Strategy drew our attention to schools, which had been closed, or were duplications.
12th May	3880	3883	Three (non KS3 Maths schools) who weren't previously on our radar, have started accreditation.
19th June	3883	3884	One (non KS3 Maths schools) who weren't previously on our radar, have started accreditation
7th July	3884	3885	Two (non KS3 Maths schools) who weren't previously on our radar, have started accreditation
			Two schools have merged. URN 105356 is now redundant.
21st July	3885	3887	Two who weren't previously on our radar, have started accreditation
4th August	3887	3878	"Closed" schools were identified by the red school Strategy and removed
1st September	3878	3877	Becta identified a school had closed (URN 108099)

Table 10: List of changes to total number of eligible schools

268. Given that the number of schools to be counted is potentially variable, and that schools with very few suitable pupils can affect the count of how many schools are participating quite significantly, consideration is being given to changing the unit to be counted for participation purposes from 'schools' to 'pupils'.

5.3.1.2 Tracking status of schools

269. The programme has categorised schools into four colours, according to their participation status:

- **Red**: school that has never expressed an interest or informed RM they are withdrawing from the test altogether. Or a school that has expressed an interest but not provided adequate contact details
- **Orange**: school that has expressed interest, but either has not started the network audit, or is in the process of the audit. Also includes schools who fail on technical grounds.
- **Yellow**: school that has successfully passed the network audit but has not installed the software.
- **Green**: school that is accredited and has installed the test software – may or may not have participated in a pilot.

© 2006 QCA

270. Each colour of school is the responsibility of a different agency, as shown in the following figure:

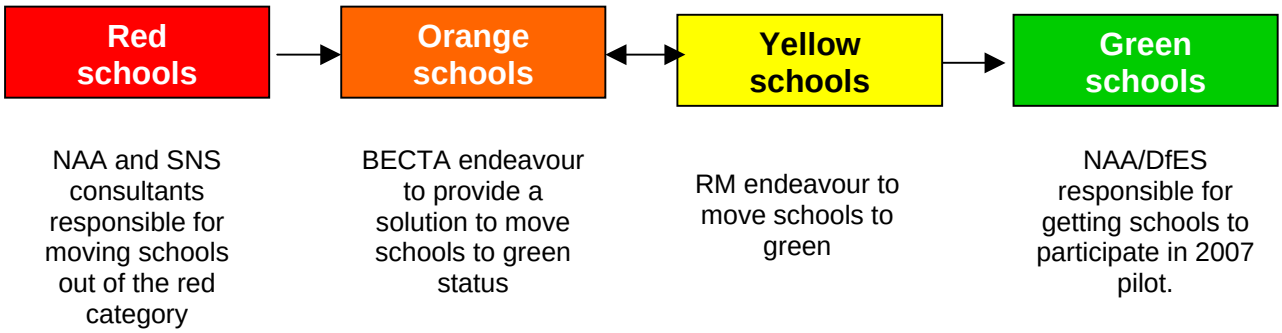

Figure 5: Responsibilities for ensuring participation in KS3 ICT test pilots

271. The colour categorisation provides the clearest way to demonstrate schools' participation status. The following table and figure show the status of schools at 26th May 2006 (that is, the last day of the 2006 pilot):

R	Haven't expressed an interest	509
	Withdrawn	90
	Expressed an interest not provided further contact details	73
		672
O	Not started audit	195
	Running audit	73
	Running audit but found a technical problem	72
	Field trial with no software	28
	Failed audit	34
		402
Y	Passed audit, not agreed terms and conditions	65
	Awaiting software	1
	Field Trial with old software	12
	Received software not yet installed	237
	Installed software – but not patch	233
		548
G	Installed software and patch	2152
	Field trial with patch	112
		2264
	Total	**3886**

Table 11: Participation status of schools as at 26th May 2006

© 2006 QCA

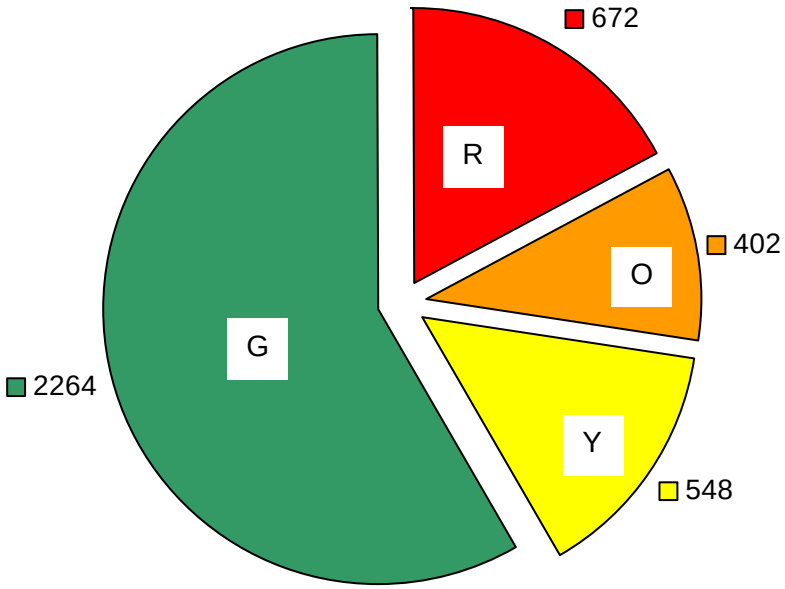

Figure 6: Participation status of schools as at 26th May 2006

272. It is appropriate to repeat the critical success factor associated with objective six:

> fewer than 50 maintained schools fail to register interest; fewer than 50 fail the technical accreditation; and fewer than 200 complete accreditation but do not take part in the 2006 summative test window (for the avoidance of doubt, this permits 300 or fewer maintained schools not participating in 2006)

273. Comparing the results in the table and figure above with the CSF requirements, one can see:

- 509 schools did not express an interest in the test (rather than 'fewer than 50' required by the CSF).
- 402 schools were in the 'orange' category – rather than the 'fewer than 50 fail the technical accreditation'[9].
- There were 548 schools in the yellow category – rather than fewer than 200 completing accreditation but not taking part.

274. It is also worth comparing the figure of 2264 schools in the 'green' category on 26th May with the 1762 schools that sent valid test data back from the pilot.

[9] The CSF states that fewer than 50 will fail technical accreditation. 47 schools had attempted to carry out technical accreditation and had failed at 26th May. The figure of 402 schools includes schools that had neither started nor completed technical accreditation by the end of the summative test window.

© 2006 QCA

5.3.1.3 Evaluation of participation

275. The comparisons above make it plain that in the 2006 pilot insufficient schools took part for the CSF to have been complied with. The most natural interpretation of that finding is that the pilot has not achieved this facet of objective six.

276. However, the CSF to objective six has a second clause. That is:

> if these targets are not achieved, a credible action plan is in place to ensure targets are achieved in the 2006/07 school year.

277. Plans for achieving full participation in 2006/07 are discussed at paras 310ff.

5.3.2 Monitoring of schools' experience

5.3.2.1 Questionnaire findings

278. The test development project's 'service delivery report' (written by RM and accepted by NAA project staff) reports the opinions of school staff at different stages in the participation process.

279. The responses to 191 questionnaires, which had questions about the following issues, were analysed:

- Expressing interest
- Network audit
- Terms and conditions
- KS3 ICT website
- Use of customer services team

280. The response profile was very good, with approximately 94 per cent of respondents agreeing (or strongly agreeing) with positive statements put to them.

281. Particular areas of strength included:

- the clarity of instructions for the expressing interest and network audit
- the clarity of information from the customer services team

282. Areas where respondents were less happy (and where, hence, lessons could be learned) included:

- Sufficiency of information provided within frequently-asked questions (FAQs)
- Ease with which to find information on the website
- Time taken to carry out the network audit

283. 230 schools returned questionnaires relating to the installation of the infrastructure and test software.

© 2006 QCA

284. The results for this stage of the participation process were good, although not as outstanding as those for the accreditation process. Particular strengths included:

- Satisfaction with technical support services (this had several facets)
- Ease of creating users on the APS

285. However, there were other areas with less favourable response rates. When asked whether they would recommend to other schools that they take part in a future pilot, 170 schools said that they would do so, whereas 47 said that they would not (a ratio of 78%:22%).

286. A series of positive statements about the installation process had an agreement rate (including 'strongly agree') of approximately 82 per cent. Similarly, about 82 per cent of respondents thought that the amount of information in the various installation guides was 'just enough'. Approximately 86 per cent of respondents rated the various installation guides as 'easy' or 'very easy' to navigate.

287. Analysis of feedback questionnaires from a small sample (60 schools) on guidance material used during the May test window was reported.

- 73 per cent of respondents believed that the guides used to assist in running the test during the summative window made it easy or very easy to locate information.
- 89 per cent of respondents agreed or strongly agreed that the guides contained sufficient information to facilitate running test sessions.

288. A key lesson that the project feels it has learned from the 2006 pilot is that it will need to improve the indexing of written guides for future years.

289. Schools' feedback on their experience of technical support and customer services were reported.

- 98.5 per cent of respondents understood the information they received from the technical support team.
- 93.8 per cent of respondents were happy with the quality of the team's response to the technical support query.
- 91.7 were happy with the speed with which the team dealt with the technical query.

290. Findings with respect to customer services were as follows:

- 99.3 per cent of responding schools understood the information they were given.
- 94.6 per cent were happy with the quality of the team's response.

© 2006 QCA

5.3.2.2 Reasons given for not participating

291. Questionnaire responses from schools participating in the 2006 pilot give a useful indication of the opinions about the test of all schools. However, there is a substantial proportion of eligible schools in England that is not yet participating in the implementation of the test. It is important to attempt to establish the opinions of such schools, in order to have a fuller understanding of the reception that the test is likely to receive across the country.

292. Establishing the attitude towards an innovation of non-participants is very problematic[10]. This is basically because of a lack of information; if people do not take an interest in an innovation, it may well be that they say nothing about the reasons for their lack of interest. Further, when they do give a reason for non-participation, it may be wise for a researcher to interpret this reason cautiously; non-participants may be reluctant to fully describe the (lack of) motivation which underlies their decision to abstain.

293. At the time of writing a solution to this methodological dilemma has not been found, but the dilemma should be borne in mind when interpreting findings below.

294. There are two sources of information into schools' lack of participation in the 2006 pilot. Firstly, there are sets of emails and notes of phone conversations explaining why schools withdrew from the 2006 pilot. Also, there is a set of observations from Secondary National Strategy consultants categorising the reasons for schools' non-participation. These two sources of information are reported below.

5.3.2.2.1 Reasons for withdrawal given in emails

295. A spreadsheet has been provided, containing the reasons why 102 schools withdrew from the 2006 pilot. These reasons were conveyed to RM by school staff in extracts from emails, or are contained in summaries of phone conversations.

296. The set of reasons seems much more homogenous than a similar list of reasons for non-participation which was analysed in the 2005 evaluation report.

297. There are several interpretations of the reasons for why schools did not proceed with the test in 2006. An informal summary of the 2006 reasons for withdrawal can support the view that the largest category is of schools who prefer

[10] There is a similar issue with respect to the seeming lack of uptake of formative reports: see para 236.

© 2006 QCA

to 'wait and see' how the test develops, rather than be proactive and participate in the 2006 pilot[11].

298. Examples of 'wait and see' type responses include:

- We will be pleased to receive the materials, but will not be taking part in the actual pilot. We will use the materials to verify our teacher assessments.
- With respect to the pilot exams, would it be possible to do a pilot examination in 2007 only, as opposed to 2006 and 2007?
- Unfortunately we will not be in a position to support the pilot test this year. We look forward to seeing the outcomes of the pilot and plans for implementation in 2007.

299. As well as the group of schools that appears to want to wait and see, there are small groups of schools with miscellaneous reasons for not participating. Such reasons include:

- staff shortages
- technical problems (generally admissions of inadequacies with the school's infrastructure, *not* complaints about the technical quality of the test software)
- explanations that the school is 'non-standard' in some way (e.g. very small, a special school, a PRU, etc.).

300. It is notable that there is a relative absence – when compared with 2005 – of schools criticising the ICT test (on educational grounds) or of schools saying that they had tried to implement the pilot, but had withdrawn due to logistical reasons (e.g. unreasonable burdens[12]).

5.3.2.2.2 Strategy consultants' categorisation of non-participating schools

301. A piece of work was undertaken in which NAA staff concerned with participation and Wider-School Readiness (see para 313 below) developed a set of categories to describe red schools. There were five main categories:

- Leadership
- ICT staff
- Technical
- Pupils
- Establishment

[11] This is one way of grouping the stated reasons for not proceeding, and different interpretations could be made.

[12] As in any understanding of people's reasons for not doing something, this interpretation could challenged. For instance, if probed, those schools citing time as the reason for not being able to take part may have felt that the software placed an 'unreasonable burden' on the school which the staff did not have the time to carry out.

© 2006 QCA

302. Each category was then sub-divided to permit a more fine-grained description of the reasons for schools' non-participation.

303. Secondary National Strategy consultants visited schools and assigned categories to each school. The following table summarises these categorisations:

© 2006 QCA

Grouped category	Reason for red status	Total
Establishment	This is a fragile establishment (e.g. special measures) and unable to cope with multiple priorities.	24
	Other reason associated with the establishment.	36
	Sub-total	**60**
ICT staff	There is a significant ICT staff shortage due to unfilled vacancy, maternity leave, sickness or similar.	34
	The ICT staff teach a different qualification (e.g. DIDA).	12
	Although there are adequate ICT staff they are too busy and have other priorities.	71
	The curriculum is taught in a way that does not 'suit' the test.	8
	The ICT staff do not like the test.	4
	Other reason associated with the ICT teaching staff.	17
	Sub-total	**146**
Leadership	The senior leadership is not supportive and has encouraged staff not to participate.	48
	Other reason associated with senior leadership.	19
	Sub-total	**67**
Other	Another reason not explicable by any of the above statements.	22
	Sub-total	**22**
Pupils	They have no pupils (or a tiny number of pupils) in year 8 or 9 working at level 3 or above.	363
	They have relevant pupils that they do not believe are ready for the test.	10
	Their pupils are working on other priorities (non-ICT) as determined by the school.	7
	Other reason associated with the pupils at this school.	11
	Sub-total	**391**
Technical	They have experienced technical problems with the test software or other related RM materials.	4
	They have a Windows infrastructure but believe – or know – that they will not meet the minimum specification.	10
	They have a non-Windows infrastructure and are awaiting further guidance/developments.	10
	They are awaiting new equipment (in under 6 months) before further activity.	18
	They are awaiting new equipment (in over 6 months) before further activity.	2
	They have suitable equipment but there are other priorities on the network.	1
	Other reason associated with the technical infrastructure	10
	Sub-total	**55**
Other	Another reason not explicable by any of the above statements.	22
	Sub-total	**22**
#N/A		3049
	Don't know and it's difficult to find out.	52
	No information available at this time	32
	School falls outside the jurisdiction of the LA (e.g. Academy)	11
	Sub-total	**3144**
	Grand total	**3885**

Table 12: SNS consultants' categorisations of reasons for non-participation

304. Several rows from this table support findings that have emerged elsewhere in this section of the report:

© 2006 QCA

- Within the 'establishment' category, there are 36 schools where there is '[an]other reason associated with the establishment'. When further explained many of the 'other reasons' describe non-standard schools', such as special schools, PRUs, hospital schools, etc. This confirms the impression that such schools might be a special case for participation purposes.
- Within the ICT staff category, there are 34 schools where there is a significant ICT staff shortage. This is consistent with schools' own reports when giving reasons for quitting the pilot (see para 299 above).
- There is a very large group of schools (363) who are reported to have no or very few pupils working at the right level for the test. Further study of these schools shows them to be mostly PRUs or special schools.

305. In addition to those rows of the table which confirm other findings on participation, there are several rows which suggest findings that were not obvious from other activities reported in this section of the report:

- 24 schools are described as being 'fragile establishments' (for example schools in special measures). It seems that such schools are more likely to not participate than other schools with fewer pressures across the board.
- It seems, on the face of it, a source of concern that there is a relatively large group of schools (71) where 'although there are adequate ICT staff they are too busy and have other priorities'. Unfortunately, it is not possible to describe any other common features of this group of schools from the information currently available.
- Within the 'leadership' category, there is a substantial group of schools (48) where the Senior Leadership are either not supportive, or are reported to have actively discouraged ICT staff from participating in pilots. This is a discomforting scenario, and NAA staff are making active steps to remedy situations where necessary support from SLTs has not yet been forthcoming.

5.3.2.3 Evaluation of schools' experiences

306. Thus, questionnaire findings show schools to have had a largely positive experience of the 2006 pilot. This is especially so in respect of the accreditation and software installation processes. The findings from the questionnaire regarding the test window experience remained good, but were somewhat less so than the findings on earlier stages in the process.

307. The success factors associated with objective six set a high bar; they require questionnaire findings to demonstrate that 95 per cent of respondents had a positive experience. The reported questionnaire findings approach 95 per cent approval on occasion, but do not do so universally.

308. Schools' reasons for not proceeding with the pilot are characterised mainly by an absence of any set of widespread complaints. Rather, the reasons schools

© 2006 QCA

give tend to reflect issues affecting the school which do not emanate from any perceived fault with the test or its administration.

309. Thus, the evaluation of this facet of objective six emphasises the success of the work done, but also notes that that success was not of the extremely high quality (near perfection) demanded by the success criteria. As such, this part of the objective has not been demonstrated.

5.3.3 Plans to improve participation in 2006/07

5.3.3.1 Monitoring via programme board risk register

310. The KS3 ICT programme risk and issue register contains at least five current risks that pertain to participation.

311. The following table shows – for the purposes of exemplification – an edited portion of two of these risks:

© 2006 QCA

RISK DESCRIPTION		RISK ASSESSMENT		PRE-WARNING SIGNS	EXISTING CONTROLS	COUNTER MEASURES
CAUSE	EFFECT	IMPACT	LIKELIHOOD			
Some schools will not have adequate ICT infrastructure (hardware, software & netware) to be able to administer trials, pilots or tests during 2007 - 2008	Shortfall in number of schools taking part in final pilot. Some schools not adequately prepared for statutory test. Could fail to achieve criteria for going statutory.	SIGNIFICANT	CERTAIN	List of schools claiming to have insufficient funding, SNS checks, non-participating schools in the 2007 pilot	Funds have been made available to schools to purchase ICT infrastructure through Standards Fund and capital allocations. Developed clear goals and targets for KS3 ICT for school accreditation Fortnightly monitoring against agreed targets RM and Becta logging calls from schools who have failed accreditation (allows for targeted support) RM's fortnightly report on school participation Weekly report on accreditation Contact database of schools (Becta) Red school strategy implementation plan DfES has been provided with a list of schools that have insufficient funding, so that targeted communications can be sent to these schools	Communications sent to schools to ensure they are aware that imminent statutory status of test requires them to take action Becta support for schools which fail the “health check” DfES to write to schools who claim they do not have the funds to acquire the appropriate equipment Schools use accredited LAs to run the test (i.e. provide an alternative test site) Becta to follow up the 30 red schools who are waiting for new equipment and/or are believed to have technical problems and to identify any with lack of funding. Approach to be integrated with NAA Contact Strategy
Schools, Teachers and NAHTs are reluctant to take part in national pilots.	Some schools may campaign against the pilots. Insufficient numbers for the pilot. Higher risk of not meeting statutory requirements.	SIGNIFICANT	POSSIBLE	SNS reports Chat room feedback Low participation in 2006	High level crisis management plan has been completed, which provides lines to take so that negative publicity can be countered quickly and robustly NAA stakeholder group is in place (consisting of Teacher Union Groups) to gain feedback and keep teachers in the loop Regular SNS visits to schools Red and Amber School Strategy to target non-participation Becta's work with schools failing the Audit Contact strategy with comms (schools) plan developed to address the varying groups and apply targeted communications and strategy towards each group	Communication to be sent to schools to encourage participation through clear upfront messages about the purpose, goals and benefits of the test and risks of not being well-prepared for introduction of the statutory test in 2008 Good Practice Guide to be sent to Schools which will document case studies that show the positive attitude taken by school staff towards the test. Individual discontented schools to be contacted by consultants so that they can take swift action Proposal to meet the Head of Academies Network to discuss action to be taken to ensure academies are engaged in the ICT Test to be discussed. DfES to send out a letter to the Directors of Children Services to highlight the shortcomings in LA areas and provide a list of school status Newsletter to Senior Leadership Team to keep the pilot and test on the SLT radar

Table 13: Examples from the KS3 assessment programme board risk register

© 2006 QCA

312. The presence of these five items on the risk register illustrates the following:

- That the most senior body responsible for KS3 ICT assessment has identified non-participation as a major risk to the successful implementation of the test by 2008.
- That substantial activity is already underway to attempt to achieve full participation for 2008.
- that large strands of new work are being planned to further positively influence participation.

5.3.3.2 Wider school readiness project

313. Following the transfer of programme management to NAA (see para 375 below), a new strand of work was set up as the 'Wider School Readiness' project.

314. The premise behind WSR included an acknowledgement of the following challenges faced by schools:

- The test will challenge current teaching, raising the status and the stakes of teaching ICT. This challenge rightly sits with the ICT subject leader in school.
- The test also challenges the school's hardware, software and network management. A reliable, fair platform that doesn't disadvantage individual pupils isn't easy to achieve. Accrediting the network and ensuring access to computers that function well and cope with the required workload are part of the responsibilities of the school's technical staff, typically the Network Manager or senior technician.
- As in most schools there are more pupils than computers and up to 75 minutes needs to be allocated to run each test session (i.e. longer than a teaching period), it will be necessary to create a comprehensive schedule. This will require staff to co-operate and plan ahead to use resources efficiently. The devil will be in the detail of ensuring that all those affected know what is happening when and to whom. This detailed work is typically the responsibility of the school's Exams Officer. This role also encompasses using the Pupil Manager and Test Manager functions of the test software.
- This all takes place at a time when other national curriculum tests are taking place and there's also a significant demand for resources for GCSE/GCE exams, coursework or revision. These resources such as time, equipment, accommodation and teaching will need to be negotiated 'in competition' with other subjects, exams and tests. Ensuring the right level of priority is achieved is the responsibility of someone from the Senior Leadership Team.

315. Further, the WSR project was conceived of as filling a 'gap' in support provision within the programme. The prior DfES school readiness project was concerned mainly with technical accreditation. In contrast, the NAA-based WSR project focuses on the fact that schools that become technically accredited do not necessarily go on to participate in a pilot.

© 2006 QCA

316. The following figure was included in the executive summary of the WSR proposal. It shows the alleged gap in the provision of support to schools, and hence formed part of the justification for the instigation of the WSR project:

© 2006 QCA

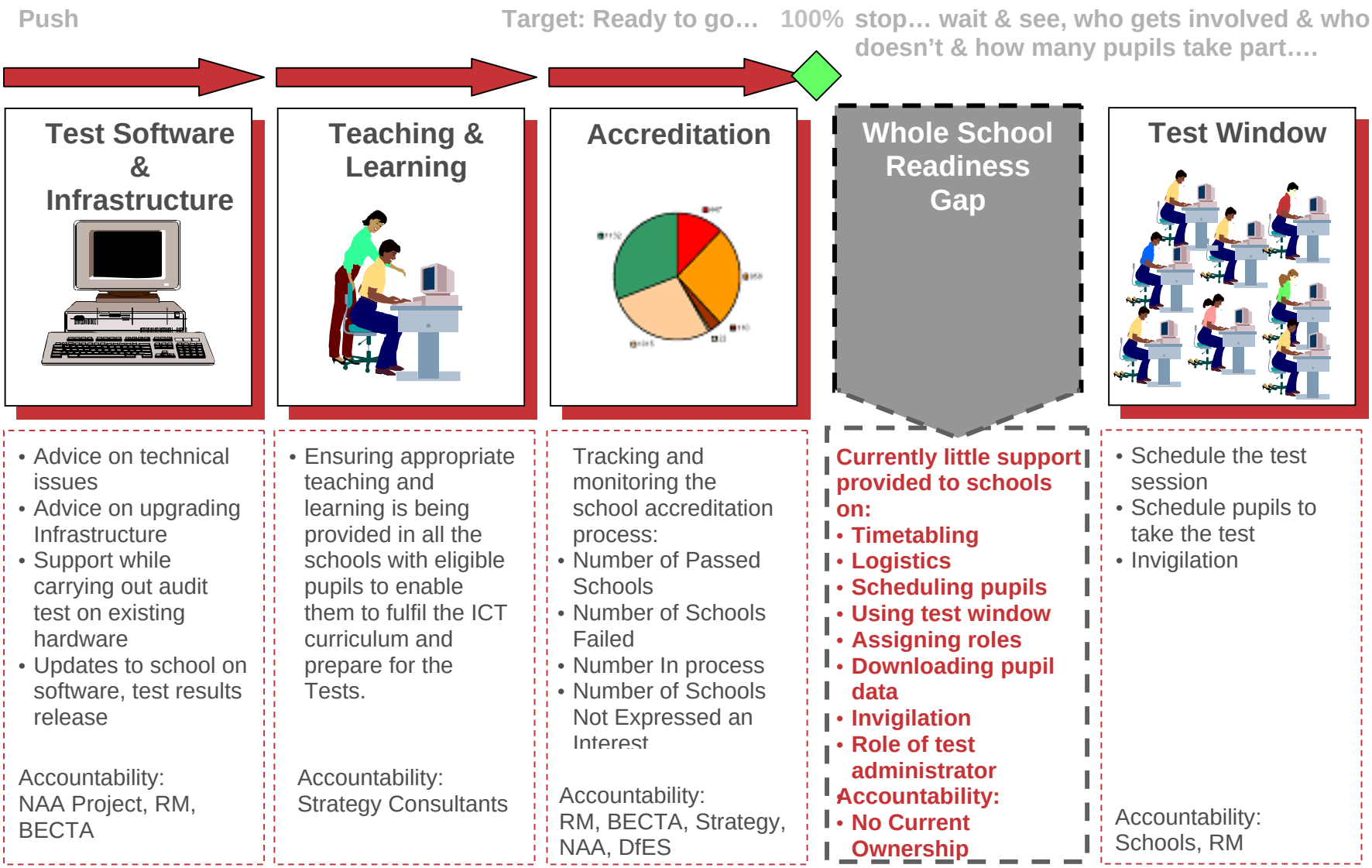

Figure 7: The 'wider-school readiness gap'

© 2006 QCA

317. Thus, the aim of WSR and other strands of the programme is to achieve the participation of 100 per cent of appropriate pupils in 100 per cent of eligible schools in the 2007 pilot.

318. To achieve this end, the WSR project is proposing three main initiatives:

- An Essential Guide
- An online version of the EG, managed by NAA webmasters, and updated throughout the duration of the 2006/07 pilot
- Increased support for NAA Field Support Officers (FSOs). In particular, it is the intention that FSOs will enable school Exams Officers to engage more effectively with administrative duties necessary to facilitate delivery of the KS3 ICT test.

319. The EG is designed to be a collection of ideas about 'what works' and positive experiences from test administrators. It offers advice on how to schedule and administer the tests, and guidance concerning roles and responsibilities.

320. The content of the EG is based on a data collection from case study visits to 22 schools.

321. In addition to the products described above, the programme's integrated communications (comms) strategy has been developed. An important element of this strategy is the allocation of responsibility for contacting schools to the best-placed organisation. This allocation is based upon a division of schools into colour groupings, based on their readiness status (see para 269 above).

322. The allocation of colour-grouped schools is as follows:

- Red schools – Local Authority and Secondary National Strategy ICT consultants
- Orange schools – Becta
- Yellow – RM, by sending installation reminder emails
- Green – FSOs supporting EOs

323. The engagement (or otherwise) of different actors with the varying colour-grouped schools is believed to be crucial to full participation in 2007, and hence to the fulfilment or otherwise of the WSR's key aim for next year.

324. It is understood that the WSR will contribute fully to the statutory readiness plan (cf. para 15 above) and will – in that way – provide input into the decision as to whether to go for a full statutory roll out in 2008.

325. It is further understood that many of the WSR's activities in the 2006/07 pilot are envisaged to be transferable to operational NAA work groups assuming a decision to move to statutory delivery following the 2007 pilot.

© 2006 QCA

5.3.3.3 Early indications of 2007 participation

326. The participation of schools is continually tracked by the programme and updated on a weekly basis. The most recently available information at the time of writing is from September 1st 2006. Whilst such information could be considered beyond the scope of this evaluation – since it relates – strictly speaking – to the 2007 pilot, it is considered relevant in that it allows the report to understand the extent to which plans for achieving 100 per cent participation in 2007 are credible – that being the second aspect of the CSF pursuant to objective six (see para 260 above).

327. A useful addition to recent tracking information is a set of graphs showing how many schools are progressing through participation stages, and comparing this information to targets.

328. Figure 8 shows progress against the target for reducing the number of 'red schools', and Figure 9 shows the progress against 'orange' school targets.

© 2006 QCA

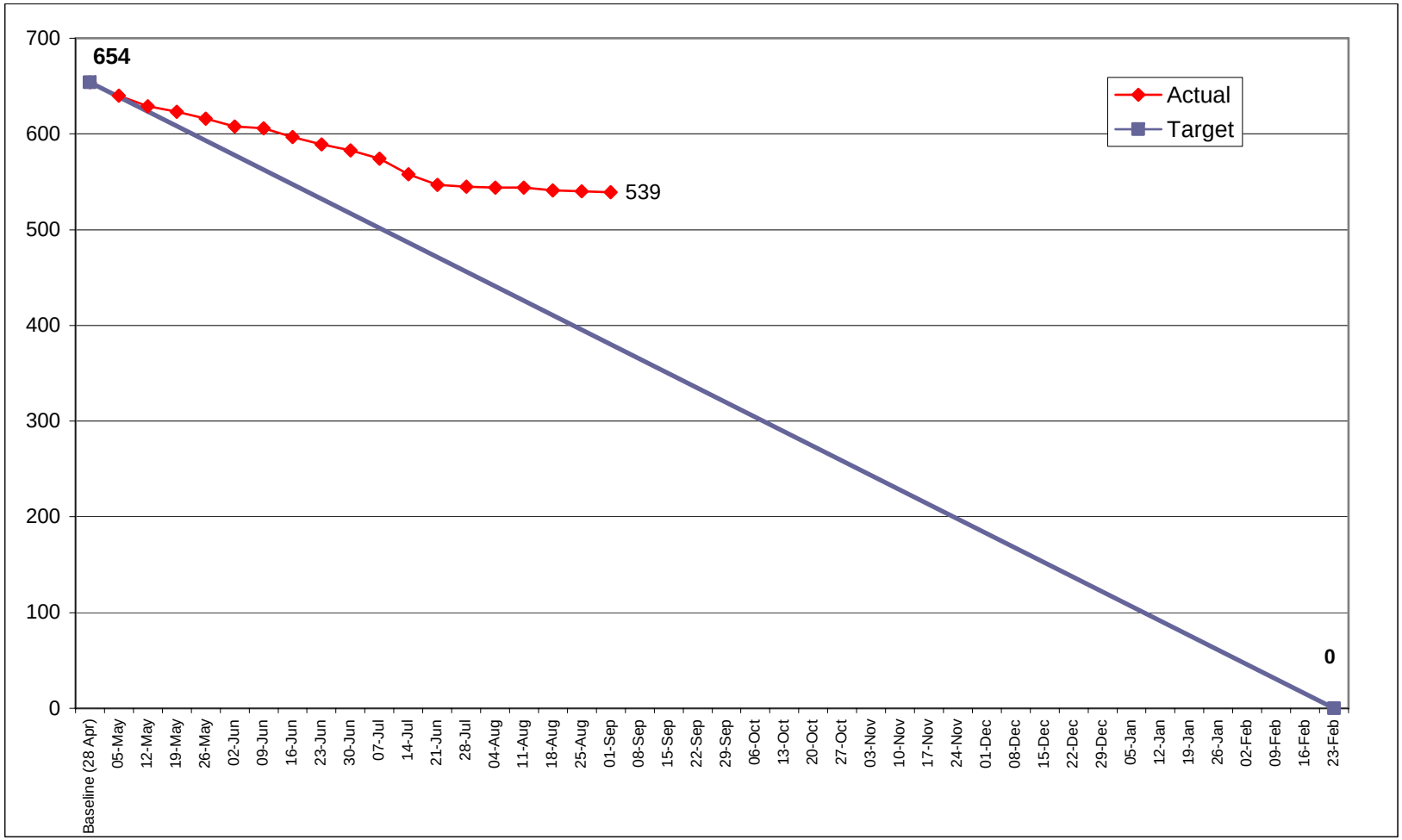

Figure 8: Target and actual numbers of red schools

© 2006 QCA

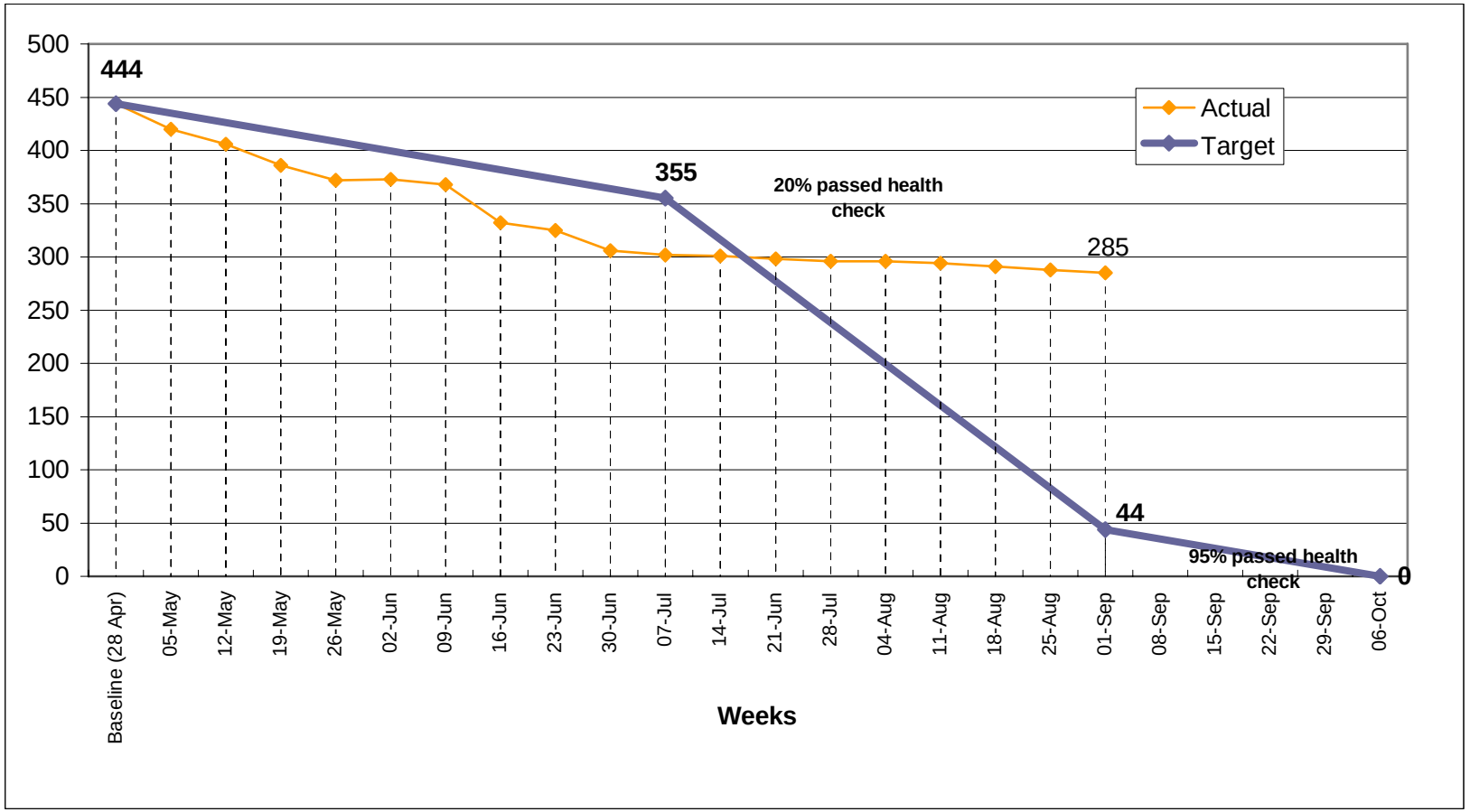

Figure 9: Target and actual numbers of orange schools

© 2006 QCA

329. Taken jointly, the two figures show that – so far – participation targets for the 2007 pilot are not being met. A particular concern is that Becta's hopes that network mangers and ICT technicians would be able – and/or inclined – to use the otherwise lighter working period of the school holidays to deal with outstanding technical issues have not been realised.

5.3.3.4 Current tactics to increase 2007 participation

330. Recent documents show that specific plans are in place to increase participation in the 2007 pilot. These plans are organised according to the participation colour category of the schools.

331. Red schools (being the furthest category from participation and full readiness) have a particular focus. A two-pronged approach is being carried out to target these schools. The DfES is targeting Local Authorities to encourage them to take action, and schools are being engaged via SNS consultants and LA strategy managers.

332. Some specific tactics aimed at increasing red-school participation include:

- Schools categorised as having leadership, curriculum or staff issues (cf. Table 12, on page 69) will be engaged by SNS consultants and LA strategy managers. This engagement will take several forms, including initially running seminars in the autumn of 2006, with non-attending schools followed up.
- NAA is about to telephone all small schools (e.g. PRUs, hospital schools, secure units and so forth) to ascertain their particular circumstances and then we will suggest the most appropriate solution – this group of schools is now very high on NAA's priority list.
- Following the conduct of a questionnaire in the spring of 2006, Becta now has considerable information about 'orange' schools. It has sub-divided orange schools into six categories, and will take appropriate actions to facilitate each sub-category's participation – in conjunction with SNS consultants, strategy mangers and others.

5.3.3.5 Evaluation of plans to increase 2007 participation

333. This sub-section of the report has shown that the programme board has designated high-level risks in the field of participation. Following from that, a specific project to maximise wider-school readiness has been instigated. That project has detailed plans, and an obligation to produce specific products. Those products appear to be based on substantial data, and to be written in a voice that ought to communicate effectively with their readers.

334. Further, integrated communication is now being implemented, especially in the sense that groups of schools that have got stuck at particular points in the process are being targeted by the most suitable institutions or individuals.

© 2006 QCA

335. Taken jointly, all this action amounts to what the CSF describes as: 'a credible action plan … in place to ensure targets are achieved in the 2006/07 school year'.

336. Unfortunately, the early tracking figures on school participation show that the reduction in both red and orange schools is falling behind target. This finding is a substantial concern. Moreover, the concern that participation in 2007 might be less than its target should be understood in the light of the previous experience of pilots not meeting their participation targets.

337. It is for these reasons that the achievement of this facet of the objective has not been demonstrated.

5.3.4 Overall evaluation of objective six

338. Thus, three facets of objective six have been distinguished. The micro-evaluations for each facet are:

- Participation in 2006: not achieved
- Schools' experience: not demonstrated
- Plans for 2007: not demonstrated

339. Taking the overall evaluation of objective six as a summation of these three micro-evaluations, the finding is that this objective has not yet been demonstrated.

© 2006 QCA

6 Concluding remarks

340. At the start of this report, it was stated that it was not appropriate to make an overall evaluation of the KS3 ICT assessment against its principal aims. This was because the report does not consider all facets of the programme (for example it is silent on software robustness).

341. However, it is felt appropriate to conclude the report with a few remarks giving the evaluator's view of key issues for the programme at this juncture. Such remarks go beyond reporting the plain evidence derived from the pilot; the robustness of the remarks is improved however, by their being firmly based in the evidence that the evaluator has observed, and by being vetted by several senior colleagues.

6.1 Quality of work in the 2006 pilot

342. The evaluation has judged the 2006 pilot not to have demonstrated achievement of the three objectives that it considered. That could be taken as an indication that the pilot had not succeeded overall, or that the standard of work produced during it was less than optimal.

343. Such inferences should only be drawn cautiously; the evaluator considers that a large amount of high-quality evidence has been made available to facilitate an understanding of the state of the ICT assessment programme. Clearly-written, precise technical reports have been provided by the test development project; 2006 information was more straightforward to interpret than in previous years.

344. And yet, the 2006 pilot has not demonstrated achievement of its three evaluated objectives. To provide some explanation of why achievement of the 2006 objectives was challenging, the notion is advanced that there are intrinsic tensions observable when considering the KS3 ICT test. These are described in turn in the section below.

6.1.1 Tensions implicit within KS3 ICT

6.1.1.1 Robust summative test vs. innovative test

345. The DfES' letter to the QCA (see para 361 below) gives the KS3 ICT test two purposes:

- A robust instrument capable of providing data for high-stakes summative purposes.
- An innovative instrument capable of showing how ICT-based assessment might best be developed.

© 2006 QCA

346. Clearly, there is a tension between these two aims. If the sole aim were to create a robust measure to indicate pupils' national curriculum level, it is arguable that a much simpler test would have been the most appropriate instrument. The converse is also true; the need to produce robust data to support accountability purposes has limited the amount of innovation that has informed this test.

347. This tension was arguably present in the original remit that defined the programme.

6.1.1.2 Novel validity questions in a conservative environment

348. The fact that the KS3 ICT test design is unlike any NC test (or – to the knowledge of this writer – high-stakes public examination) that has previously been developed in England means that the methods that have been relied upon to demonstrate the validity of existing tests are not wholly applicable to KS3 ICT.

349. This phenomenon can be observed in several areas, for example:

- Reliability (especially the need to rely primarily on classification consistency as a measure of reliability)
- Concurrent validity where there is no test that is even slightly similar
- The use of opportunities rather than marks

350. It must be emphasised that the responsibility for the state of the KS3 ICT test model lies with the test development project; that is, with the test development agency that originated the model, QCA, NAA and other stakeholders who approved work and provided advice. To the extent that some parts of the model remain unproven or even inchoate, the responsibility needs to be clearly acknowledged.

351. However, whilst it was the test development project's responsibility to choose an effective assessment model and validation methods, the obligation to develop an innovative e-assessment implicitly required any developer (and evaluating researchers) to go outside the comfort zone offered by traditional techniques.

352. Such a move outside the comfort zone must – in the short-term at least – make it more of a challenge to both design and validate the test.

6.1.1.3 Supportive yet demanding evaluation

353. The evaluation for the test development project was commissioned by that project. The intention was to provide helpful and immediate feedback to allow the production of a better test.

354. However, the evaluation was also set up to provide a certain degree of (although not total) independence from the project. In that way, the intention was

© 2006 QCA

that the evaluation would be an independent source of information for stakeholders in the programme; both to reassure such stakeholders wherever possible, and to flag concerns where that was appropriate.

355. The demanding standards required by the evaluation reflect two further facts. First of all, in recent years the science of assessment validation has moved on considerably. Documents like the American *Standards for Educational and Psychological Testing* have described the obligation on test 'sponsors' to provide clear evidence of a test's validity; it is for those putting forward a new test to demonstrate validity to potential users, rather than for commentators to 'prove invalidity'.

356. The second point is that national curriculum tests have been around for a number of years, and – in that time – have been criticised from a range of directions. Given that, it is only right and proper that the introduction of a new NC test is supported by the most robust validity evidence that can reasonably be provided.

6.1.1.4 A risky test in a risk-averse environment

357. Several of the points made above are consistent with the view that – although a large amount of high-quality work has been undertaken to establish the quality of this test, that quality cannot yet be unequivocally asserted.

358. In the light of the timescale for statutory roll out, this is clearly a substantial risk. Further, it is a risk in a government-backed ICT initiative, and – simultaneously – in a national curriculum test. Thus, the probability that less-than-optimal implementation would attract unfavourable public comment seems quite high.

359. Given that risk, it might be tempting for decision makers to attempt to reduce risk by de-scoping the project in some way (for example by requiring a more 'traditional' test design to be implemented in order to provide a more straightforward route for gathering robust accountability data).

360. In the view of this evaluation, that temptation should be avoided. Rather, the more difficult path that has been pursued to date should be continued; that is, the innovative model of test should be reaffirmed, and clear evidence of its appropriateness should be demanded before it can be used for statutory purposes with key stage 3 children.

© 2006 QCA

7 Appendix A: Background to the KS3 ICT programme and project

7.1 Aims of the programme and project

361. The DfES wrote to the QCA on 15th September 2005, and described the purpose of the assessment of information and communication technology at key stage 3 programme as:

> to develop an on-screen test to provide an independent measure of ICT attainment at key stage 3 and to indicate the potential for the wider use of electronic testing in national curriculum tests and public examinations.

362. In the same letter, the DfES reaffirmed its commitment to administering the test on a high-stakes basis with school-level results data published in 2008, subject to a successful pilot of the test in 2007.

363. School results data published in 2008 would permit the evaluation of a Public Service Agreement (PSA) target. That target would be for 85 per cent of test-taking pupils to have achieved level 5 or above in the test.

364. For that full rollout to be achieved it will be necessary that, in addition to a fully-proven test, schools have the necessary infrastructure and have undertaken adequate preparation for the administration of the new style of test.

365. 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. (*QCA Project Initiation Document, dated: 13th March 2004*).

7.2 Project contractor

366. QCA's prime contractor – responsible for overall delivery – is Research Machines PLC (RM); a provider of ICT software, services and infrastructure to UK educational institutions. RM has a number of specialist sub-contractors; including, in the current context, specialists in educational measurement.

© 2006 QCA

7.3 The 2006 pilot of the key stage 3 ICT test

7.3.1 Key deliverables

367. The DfES defined the key deliverables for the 2006 pilot:

- Practice tests, to be sent to all accredited schools in the spring term preceding the summative test window
- A refined and renewed summative test to be made available to all accredited schools during a test window in the summer term
- A pupil profile, consisting of a national curriculum level and summative report, for all pupils completing the summative test (i.e. two 50-minute sessions)
- Data from the summative test to be provided in accordance with the Department's specification.

368. The precise nature of these key deliverables was defined in more detail by the project. Definition was effected through a range of specifications and other documents. An outline of important elements of the KS3 ICT test is given in Appendix B to this report (see pages 88ff).

7.3.2 Transfer of project team to NAA

369. From the inception of the project until early 2006, the QCA team running the test development project was located within a specialist development team led by Martin Ripley (latterly known as the e-strategy unit).

370. In February 2006 the KS3 ICT test development project was relocated to the National Assessment Agency. NAA is a subsidiary body of the QCA, responsible (amongst other things) for developing, delivering and modernising national curriculum tests.

371. Martin Ripley also left the QCA at this time. Many of Martin's responsibilities with respect to the KS3 ICT test have been taken on by Mick Walker, NAA's Director of Quality Assurance.

372. The relocation of the project to the NAA has coincided with an increased involvement in the project's activities of individuals and teams that are experienced in other NC test work; for example, the teacher panel that informed level setting and the NAA level setting meeting were observed by Regulation and Standards staff. Also, NAA's work benefited from the involvement of staff such as Colin Watson, the Head of Strategy & Policy and Mick Quinlan, the Statistical Analyst. QCA also provided enhanced research assistance to NAA's level setting activities.

© 2006 QCA

7.3.3 Delegation of programme management to NAA

373. The DfES's September 2005 letter states that it expected that QCA's Test Development Team and Department's programme strands would work in partnership to maximise the number of schools which are accredited and able to participate in the pilot tests in 2006 and 2007.

374. However, an Office of Government Commerce Gateway 0 review in October 2005 made several recommendations concerning programme governance, planning and control, amongst other things.

375. In January 2006, the DfES delegated its programme management responsibilities to the NAA. On accepting these responsibilities, the NAA established a programme team to manage the delivery of the KS3 ICT programme, address the recommendations of the OGC Gateway Review 0, and to guide the process leading to statutory delivery of the test by 2008.

© 2006 QCA

8 Appendix B: Description of the KS3 ICT tests

8.1 The assessed construct

376. The test assesses pupils' knowledge, understanding and application of key stage 3 ICT as measured against the level descriptions described in the national curriculum.

377. The construct addressed by the tests is also sometimes described as 'capability', using a definition which appears in key stage 3 national strategy ICT strand documentation, amongst other places:

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

378. The content of the test samples from the national curriculum programme of study for key stage 3.

8.2 Output measures

379. The test's principal output is a national curriculum level for each pupil. The test developers are emphatic that the test is designed to output a national curriculum level for each pupil, and not some other measure – such as a sub-level or a mark.

380. In 2006 the test can 'award' pupils a level between 3 and 6 (via two test 'tiers' – see para 390). Pupils who do not display sufficient evidence to be awarded any level are categorised as receiving level 'N'.

381. Pupils' results were sent to schools for downloading via the APS software on 30th June 2006, after level setting. As well as the single digit statement of the pupil's level, schools were sent a summative report, which was a relatively brief document which summarises what the pupil did in the summative test. The purpose of this report was to justify the NC level that pupils had been awarded.

382. If pupils complete two practice test sessions, they can receive a formative report. This document has three sections: what you were asked to do, what you did, and how to make progress. It is intended as a useful output for pupils and

© 2006 QCA

teachers to allow them to use the practice tests as a source of information to improve their learning.

383. For more information on formative and summative reports see paras 405ff and 412ff, respectively.

8.3 Test environment

384. The construct assessed in this test – ICT capability – involves 'applying technology to solve problems' (see para 377 above). As such, (and following a feasibility study and stakeholder feedback) the test was based upon interactive tasks in which pupils were required to demonstrate their capability, rather than 'static' questions in which pupils' receptive knowledge is assessed[13].

385. In order to facilitate an assessment in which pupils could demonstrate their capability a bespoke virtual software environment was developed.

386. This environment was bespoke in that it was specially designed for this test. Whilst it has many generic features in common with other software environments (e.g. location of functions on menus, common shortcut keys, etc.), it does not look or function precisely like any one specific environment.

387. The environment is virtual in that it exists only for the purposes of this test. The programs within the environment can be used as functioning applications whilst a test session is running; however, the files or other data (e.g. emails) 'created' within the environment cannot exist outside a test session (e.g. an email 'sent' by the test's email client application cannot be received by an email application 'in the real world'; a file saved to a location within the directory structure in the test environment cannot be transferred to another environment; a document created in the environment cannot be sent to a real printer).

388. The test environment contains a suite of office-style applications: a word processor, spreadsheet, database application, presentation software, and email client.

389. A file manager application allows the pupils to view a set of directories containing files that could be used to respond to tasks. Also, a web browser – complete with a search engine – allows the pupils to search the 'walled garden'/simulated world wide web to locate information, images and so on to help them to complete tasks.

[13] See para 394 for the exception to this principle.

© 2006 QCA

8.4 Test tiers and forms

390. The level 3 – 6 test was an updated version of the instrument that had been piloted in 2004 and 2005, and that for which the vast majority of pupils in 2006 were entered. The level 3 – 6 test consisted of two tiers; a 3 – 5 tier and a 4 – 6 tier. Pupils were entered into the lower or upper tier contingent upon the Initial Level Assessment (ILA) that their teacher had assigned them. (An ILA of 3 or 4 entered pupils for the lower tier; an ILA of 5, 6 or above entered them for the upper tier.)

391. There were two 'forms' of the level 3 – 6 test. This was achieved by cloning the tasks in the 'form A' by changing the context – to create a parallel form B.

8.5 Task structure

392. The level 3 – 6 tests were made up of two 50-minute sessions. Each test consisted of five tasks. Three tasks were presented to pupils in the first session, and two in the second. It was the intention of the test designers that pupils should divide their session times roughly equally between the numbers of tasks delivered in each session. Therefore, the intention was that pupils should spend about 17 minutes each on tasks 1 to 3, and 25 minutes each on tasks 4 and 5.

393. An innovation for 2006 was a 'next task' button. This permitted pupils to move on, rather than having to wait for the next task to time in (for instance, if they had finished their first task after – say – 12 minutes).

394. Another innovation in 2006 was task 1 (the first task in both tiers). Task 1 provided a series of direct questions testing the pupils' ICT knowledge skills and understanding. Response formats included multiple choice, short answer or mini-task (working on a specific problem from the PoS).

8.6 Opportunities

395. Opportunities are the smallest meaningful entity that could meaningful represent a chunk of ICT capability. Each opportunity is constructed from a series of smaller entities, which can – ultimately – be traced back to mouse clicks or keyboard inputs.

396. Opportunities are levelled; that is, each opportunity has been assigned to a national curriculum level and is treated as evidence towards awarding a level for a pupil who demonstrates it. As such, opportunities are the key entity that is counted when the level-setting process is carried out (see para 191 above).

© 2006 QCA

397. Opportunities have some properties that are different to marks; for example, each opportunity can be demonstrated or not – the absence of demonstration of an opportunity does not amount to a 'wrong answer'.

8.7 Level setting process and procedures

398. As has already been stated (paras 20ff), the 2006 test was a pilot and not a high-stakes test. An important consequence of this was that the 2006 test results had no official status as a baseline for future distributions of levels to be set to. Thus, subsequent years' tests will not be required to maintain a standard set in 2006.

399. Despite that, in 2006 a standard setting process was implemented and a national curriculum level was returned to every pupil who had completed two sessions of the level 3 – 6 test.

400. The standards-setting process was based upon the 'sufficient evidence model'; a bespoke level-setting technique developed by RM and their sub-contractors. This method essentially involved the counting of the number of levelled opportunities that each pupil had achieved. Thus, a pupil in the lower tier of the level 3 – 5 test would – for example – have the number of demonstrated opportunities that had been assigned to levels 3, 4 or 5 counted. Then, they would have the number of opportunities that had been denoted as level 4 or 5 counted, and finally the number of level 5 opportunities counted.

401. The number of counted opportunities would then be assessed against cut scores applied for each level (i.e. a cut score for opportunities levelled as 3, 4 or 5; a different cut score for level 4 or 5 opportunities and finally a cut score for level 5 opportunities). If the pupil achieved each cut score, then s/he could be awarded level 3, 4 or 5, respectively.

402. The cut scores (level thresholds) were decided upon by a level-setting meeting, chaired by NAA. Participants at the meeting included several members of NAA staff, QCA statisticians and researchers. There were also some observers at the meeting. These included QCA regulation staff and an experienced independent observer.

403. RM presented a range of information in the form of reports to assist the level-setting process.

404. The meeting decided upon a definitive set of cut scores after scrutinising suggested thresholds provided by: a teacher panel, the NAA curriculum adviser and RM's educational specialists.

© 2006 QCA

8.8 Formative reports

405. To receive a formative report, a pupil must complete a practice test. The pupil can be logged onto the practice test in: scheduled, named unscheduled and anonymous (unnamed unscheduled) modes.

406. The formative report is available at the DPS (Delivery Point System – typically the pupil's workstation) for all three modes, and is additionally sent to the APS for both named modes.

407. The formative report is produced when a pupil completes a practice test of 2 x 50-minute sessions.

408. The formative report is a plain-text document, viewable on screen (and saveable to disk) or printed off. When printed off, the report typically occupies three or four pages of A4 paper.

409. The report is divided into three sections:

- What You Were Asked To Do
- What You Did
- How to Make Progress (you need to be able to …)

410. The 'what you were asked to do' section of the report is organised to reflect the five tasks in the practice test. However, the two other sections are organised according to the four aspects of the NC programme of study for ICT[14].

411. The statements in 'what you did' and 'how to make progress' sections are based on the pupil's actions during the test. The statements are linked to triggered opportunities in the test. Algorithms are implemented in the generation of the reports to avoid repetition and illogicality.

8.9 Summative reports

412. Summative reports are returned to schools along with the level awarded for each pupil for the summative test. The purpose of the summative reports is to explain and justify the award of the particular level to teachers.

413. The summative report is shorter than the formative report, containing fewer statements from the four aspects of the NC for ICT, as well as each pupil's level.

414. In addition to the summative report, *per se*, users of the school's APS have access to various other information – such as viewing distributions of levels obtained in a summative test.

[14] That is, the following areas: finding things out; developing ideas and making things happen; exchanging and sharing information; reviewing, modifying and evaluating.

© 2006 QCA

9 Appendix C: Acknowledgements

Thanks is due to the following people who helped this report to be written.

Person	Organisation	Help provided
Sue Walton	QCA/NAA	Checking of draft report. General assistance relating to scope of evaluation, etc.
Martin Adams	RM	Provision of information in report interview. Factual checking of draft report.
Jim Brant	QCA/NAA	Provision of information in report interview.
Chris Mirner	QCA	Provision of information in report interview. Factual checking of draft report.
Colin Watson	QCA/NAA	Checking of draft report.
Vaezi Chima	QCA/NAA	Provision of information in report interview.
Jo Coutts	QCA/NAA	Provision of statistical information re participation in the 2006 pilot. Checking of draft report.
Ann Carroll	RM	Provision of information re opinion collection and analysis.
Paul Newton	QCA	Checking of draft report.
Rowena Wilkinson	RM	Factual checking of draft report.
Dennis Opposs	QCA	Checking of draft report.

© 2006 QCA