Report on
Key Stage 3 English
Review of Service Delivery Failure
2003–2004
to QCA Board

The Qualifications and Curriculum Authority (QCA) Board established the following Terms of Reference for a Committee of Review at a meeting of the QCA Board on 30 September 2004:

To inquire into and report upon the reasons for:

- the late delivery of materials to schools and to markers;
- the inadequate performance of the [e-results] website;
- the additional week being inadequate for marking to be completed.

To advise on the adequacy of the steps taken to identify and remedy these problems.

To inquire into and report upon the circumstances in which the decision was taken to return unborderlined scripts to schools.

To investigate and report upon such other matters as the Committee deems relevant to the review.

To recommend on action to be taken to avoid service failure in 2005.

Procedures

The Committee should interview appropriate personnel identified by the organisations involved, including the DfES, AQA, Pearson, NAA and the relevant [QCA] Divisions.

The Committee is chaired by Mike Beasley (QCA Board member) and includes:

Edward Gould (QCA Board member)
Sue Kirkham (QCA Board member)
Hilary Emery (DfES)

Ian Valvona (QCA Supporting Officer)

Introduction

The Committee of Review investigated the 2003–2004 key stage 3 English test service delivery failure and interviewed key individuals from Department for Education and Skills (DfES), QCA divisions, the National Assessment Agency (NAA), the Assessment and Qualifications Alliance (AQA) and Pearson. Additionally, teacher unions were invited to submit evidence to the review team and individual feedback from teachers, markers and other interested parties was sought through the QCA website.

The review team would like to thank those individuals and associations who invested their time in preparing evidence for the review.

THIS PAGE IS LEFT INTENTIONALLY BLANK

Executive Summary

The 2003–2004 key stage 3 English test operations process was plagued with myriad issues and errors. While each issue and error in itself would have been manageable, the combination of so many caused the failure.

The review team found no reason to believe that the test itself, the marking quality, or the final national results were in doubt.

The process from the printing and distribution of the test materials through to the publication of electronic results was, however, badly flawed causing significant concern and disruption in schools.

The whole process was characterised by poor leadership and inadequate project management.

In consequence the principal recommendation of this review is the establishment of a co-located team comprising seconded members of the key partners involved in delivering the process, led by a senior manager from NAA.

It is considered essential that the seconded members of this team collectively accept corporate responsibility to ensure that the test programme is delivered.

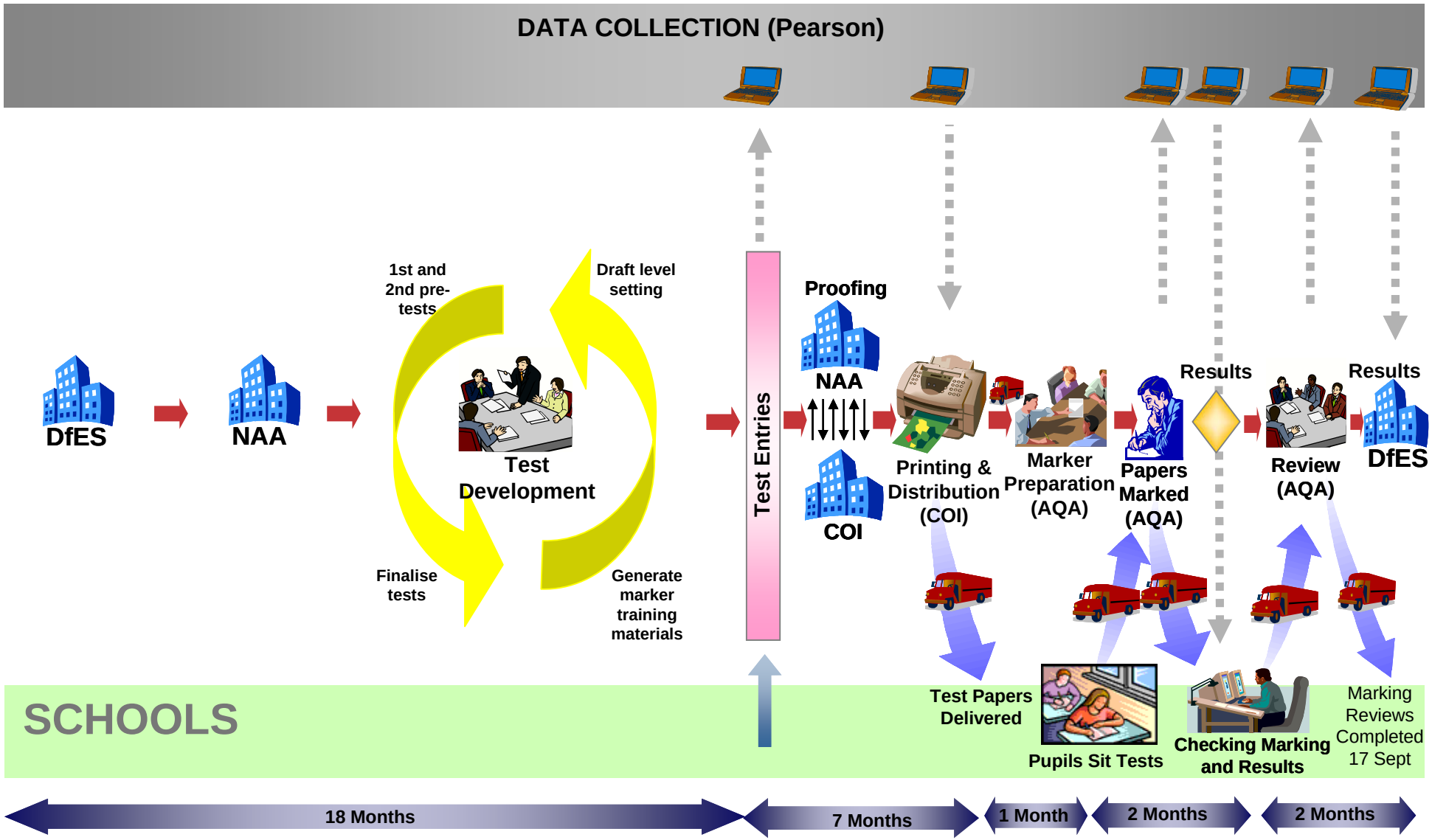
The review concluded that, whilst the issue that caused the final disruption was the publication of electronic results, the whole test operations process is not robust in any sense.

The Test Operation team's competencies in the field of process management are a major concern to the review team. It is essential that the recommendations of this report, if accepted, are executed with a thorough review of the competencies and capabilities of individual members in the Test Operations team.

Having studied the process and the time wasted through error and inefficiencies, the review team believes that, with the benefit of component marking and component borderlining, there is adequate time to complete the test delivery task to a requisite quality standard.

Figure 1

KEY STAGE 3 ENGLISH – PLANNED TOP LEVEL PROCESS (2003–04)



Background: The Key Stage 3 English Test

- Introduced as a formative and summative test regime mid-1990s.
- **2001**—Statutory targets for Year 9 pupils introduced.
- **2002**—Key stage 3 results included in secondary school performance tables.
- **2003**—Separate key stage 3 performance tables; value-added measures for secondary schools first published.
- **2004**—Date set by government for reaching first national targets for key stage 3 results.

The form of the test remained the same from the mid-1990s to 2002. In 2002–2003 the reading test changed to include more, shorter questions. The writing scores are now made up of one short and one longer writing task, with a more atomistic mark scheme than previously, designed to encourage more accuracy and precision in marking. In 2003–2004 the shorter writing task was part of the Shakespeare test.

In 2002–2003 it became possible for the first time to have separate reading and writing levels for each pupil. In 2003–2004 the arrangements for marking moved to component marking, to reduce the burden on markers and encourage more accurate marking.

> **Component marking** is the marking of writing and reading as separate parts of the English test.

The review team believes that the summative nature of the key stage 3 English test has, over time, become more significant, driven in part by government's decision to report national performance datasets. On the other hand, most teachers appear to want the tests to serve a more diagnostic, formative role. The DfES also wants the tests to serve as formative tools.

Scoping the Problem
Figure 1 illustrates the planned process.

On Thursday 6 May and Friday 7 May 2004, approximately 624,000 pupils from 4,500 state-maintained and independent schools, pupil referral units and hospitals sat their statutory key stage 3 English National Curriculum tests. Over these two days, pupils sat three separate English tests at times designated by QCA. The three tests consisted of:

- Reading paper (15 minutes reading time; 1 hour to write answers)
- Writing paper (45 minutes)
- Shakespeare paper—assesses reading and writing (1 hour, 15 minutes).

This year a very significant number of schools were affected by delays and operational difficulties.

The main problems experienced included:

Late delivery of test papers
In 2003–2004 a significant number of schools reported that test papers were delivered later than the published date which was 19–23 April 2004. On the actual test days there were still some schools that had not received their papers. Furthermore, where schools had indicated that they needed further test papers because their pupil numbers had changed, these requests were often unprocessed resulting in schools having to wait even longer for their full complement of papers.

Late delivery of administrative materials
Schools should have received all administrative materials (i.e. *Schools' Guide*, marksheets and stationery) by 23 April. QCA Customer Services received a volume of telephone calls and emails that appeared to indicate deliveries had not started until week commencing 4 May. Some schools did not have these materials before the tests took place.

Late results
Schools were advised in January 2004 that, due to new component marking arrangements, their key stage 3 results would not be available until 13 July, a week later than published in the *Assessment and Reporting Arrangements* documentation sent to all schools. Component marking meant that reading and writing marking was undertaken by two separate markers not one marker as in previous years. On 13 July only 77% of schools had received their results. The remaining schools received their results in the following few days but by this time many schools had already broken up for the summer holidays.

Incomplete results
Some schools receiving their results in July received incomplete results because, in some instances, their pupils' scripts had not been **borderlined**.

> **Borderlining** is a check that is carried out on test scripts that fall within three marks below the whole subject level thresholds. The intention of this activity is to check and adjust pupils' National Curriculum levels, if that adjustment is justified on the basis of quality and fairness.

Schools were further inconvenienced by having to return those pupils' scripts requiring borderlining, and wait until mid-to-late September for the results.

At the time this report was submitted to the QCA Board on 9 November one school (250 pupils) still had borderline marking outstanding and 18 schools (332 pupils) review marking outstanding. These scripts will be marked by 12 November.

THIS PAGE IS LEFT INTENTIONALLY BLANK

The Impact on Schools, Markers and Government

Schools

The impact of the failure on schools was clear to the review team from the earliest stage of the review. Anger, frustration and disillusionment underpinned virtually all of the submissions received from schools, and LEAs working closely with them. Those feelings were clearly linked to the increasingly high-stakes role that the key stage 3 English test plays in school life. Further exploration of this role lies outside the review team's remit but, nevertheless, it is a fact that school-level key stage 3 test results have a significant impact on schools with the potential to affect teachers' careers.

Most teachers and subject leaders work extremely hard to prepare their pupils for the key stage 3 English test. They feel that the delivery failure undermines their professionalism in the eyes of parents, to whom they need to report accurate results and impacts on pupils whose teaching and learning requirements may need reviewing in the light of the test results.

Although the NAA did apologise for the delivery failure in July 2004, there is a feeling among teachers that there has not been an appropriate recognition of the hours they have had to spend on checking papers, writing appeal letters, making phone calls chasing papers, downloading results and remarking papers. There was a perception that the marking and data collection processes had not been carried out accurately by the External Marking Agency (AQA) or the Data Collection Agency (Pearson) in the first place.

Markers

Markers have also been affected by the delivery failure. The vast majority of markers, employed by AQA, are extremely professional. Frequently, they are teachers who choose to mark pupils' scripts in order to develop professionally and gain a more sophisticated understanding of the assessment process, rather than simply to earn extra money.

From the markers' point of view the issues eroding teachers' confidence in the English test in 2004 were around data collection and results dissemination, not marking, although a perception spread that the marking process was flawed and some markers joined the public debate to defend their professional integrity. However, the review team did identify a historic inability to recruit enough markers as a key issue during the review.

It is possible that this annual difficulty in recruiting markers has implications for marking quality although the review team found it difficult to identify any evidence that clearly establishes marking quality as an issue. The criterion for judging marking quality normally used is the number of reviews submitted by schools that are upheld by the External Marking Agency (AQA), expressed as a percentage of the whole cohort. The review team found this an unreliable and unsatisfactory criterion due to the inconsistencies in the way data has been collected from year to year and the lack of information on the reasons why schools may or may not decide to request a review.

Government
The DfES sees the key stage 3 English test delivery failure as having detracted from the quest for better quality marking. Feedback to senior officials and ministers from schools and the National Strategies underscores the significant reputational damage done to the key stage 3 English tests. Publication of the provisional national statistics for key stage 3 English has been significantly delayed whilst publication of the key stage 3 Achievement and Attainment Tables has been postponed until March 2005.

Regaining Teacher and Marker Confidence
As the review team worked through the evidence it became clear that teachers and markers have lacked confidence in the key stage 3 English test for several years. This is despite a process of improvement in both the nature and approach to testing in 2003, and the fact that further issues identified by QCA were addressed in 2004. Of these, component marking was intended to address teachers' widespread concerns with marking quality.

The review team sets out in this report their view of what went wrong, why and how, in delivering the key stage 3 English test in 2003–2004, as well as recommendations to prevent a similar failure in 2005 and beyond. The 2004–2005 test must be delivered smoothly if teacher and marker confidence in the key stage 3 English test process is to be restored. The following view from a school reflects well a minimum level of service expected from a national system of assessment:

> *"As a school, we need to know well in advance what format of results will be issued to us, and by when. Final overall English levels should be clearly identifiable, reliable and accurate. The administrative burden of meticulously checking examiner marking script by script should not fall on schools and the entire process should be complete in sufficient time for us to issue accurate results to our students and their parents in advance of the end of the summer term."*

Certainly the organisations collaborating in the smooth delivery of the test in 2004–2005 should aspire to more than a minimum level of service. However, the review team is determined that future changes to the test delivery process should be fully 'stress-tested' and modelled for risk. If public confidence in the test is going to be rebuilt from 2005 onwards schools need to be able to depend on this basic level of service and quality.

Lack of Programme Management and Leadership

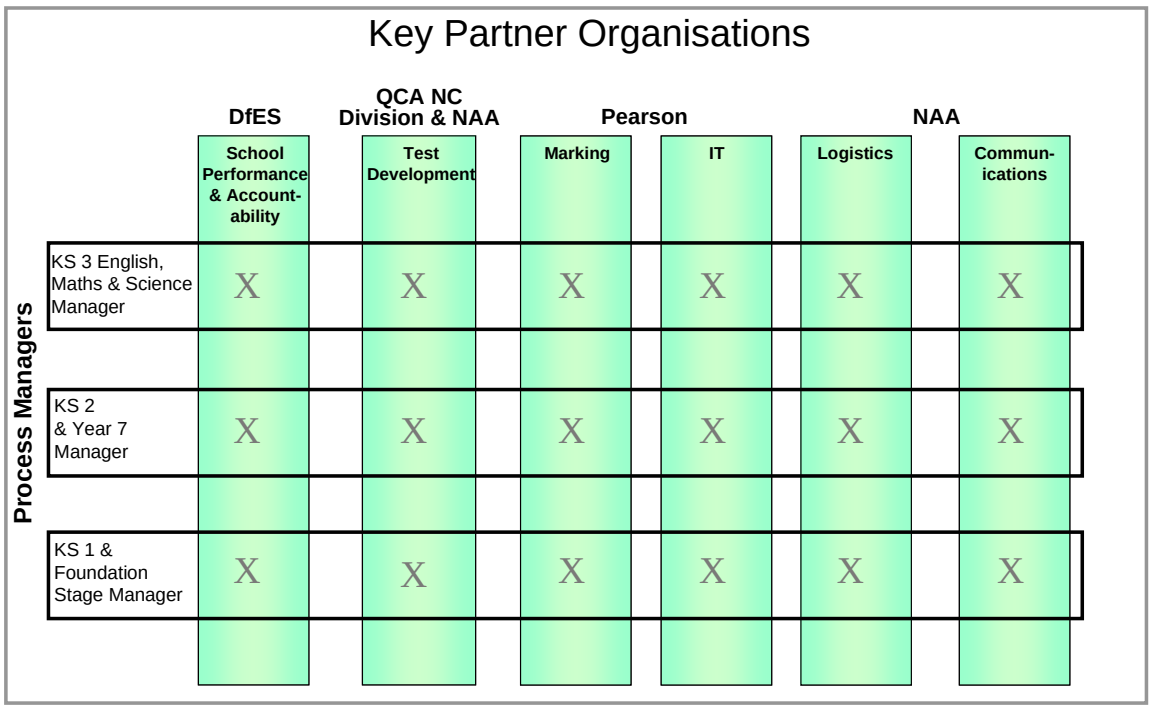
Lack of programme management and leadership were the primary root causes of the service delivery failure in 2003–2004. While there were several specific errors and poor decisions associated with this test cycle (see Annex A), the overwhelming reason for the delivery failure was a lack of effective programme and project management and an absence of overall leadership of the test development and delivery process from start to finish. This resulted in poor communications between the partners in the test process. In consequence, poor decisions and actions were taken and, as a result, the eventual delivery difficulties were unexpected and poorly managed.

Senior NAA managers do not appear to have effectively managed the arrival and departure of key personnel following the 2002–2003 test cycle and this resulted in a relatively inexperienced Test Operations team, lacking key project management competencies, delivering the 2004 test administration and marking process.

Coordination between the principal partners DfES, QCA divisions, NAA, AQA and Pearson was poor and there was no evidence of any sense of collective responsibility to achieve a positive outcome until failure was both obvious and irreversible. The operational processes used were clearly not robust. While there were many specific operational failure issues they all fundamentally stem from this lack of leadership and ineffective programme management coupled with poor communication and coordination.

The review team also heard evidence which suggested that DfES might usefully have adopted a more 'hands-on' role acting as part of the team helping to deliver the 2004 key stage 3 English test, and taking their share of ownership and accountability for it, rather than simply asking questions for information and reporting purposes, and not seeking to participate in the management or direction of the process.

Figure 2 PROPOSED NC TEST STRUCTURE 2004–05



Recommendations

Recommendation 1: in order to secure the 2004–2005 cycle, the review team strongly recommends that a matrix management structure similar to that described in Figure 2 should be established immediately.

Whilst it should be recognised that the execution of the National Curriculum test process should be a repeatable, straightforward, operational issue, a combination of historical changes and the lack of confidence by all the parties involved as a result of the 2003–2004 cycle delivery failure has caused the review team to firmly believe that the 2004–2005 cycle should be treated as a high-intensity project. Clearly the regulatory function of the QCA should not be part of this management structure.

Recommendation 2: a senior NAA representative should chair the team, provide leadership for the entire process, and all of the identified partners in the process should second key members of their organisation to the team. Subject, as well as test operations, expertise would be included.

The team would be responsible for the end-to-end process of delivering test operations. It is absolutely not envisaged that the team would set policy but would effect a rigorous change management process for any policy change requests it receives. This should be an integral part of policy development.

Seconded organisation members' responsibilities are to develop and improve on the 2003–2004 test process and to ensure that their sponsoring company and departments deliver the programme to the agreed timetable and quality standards.

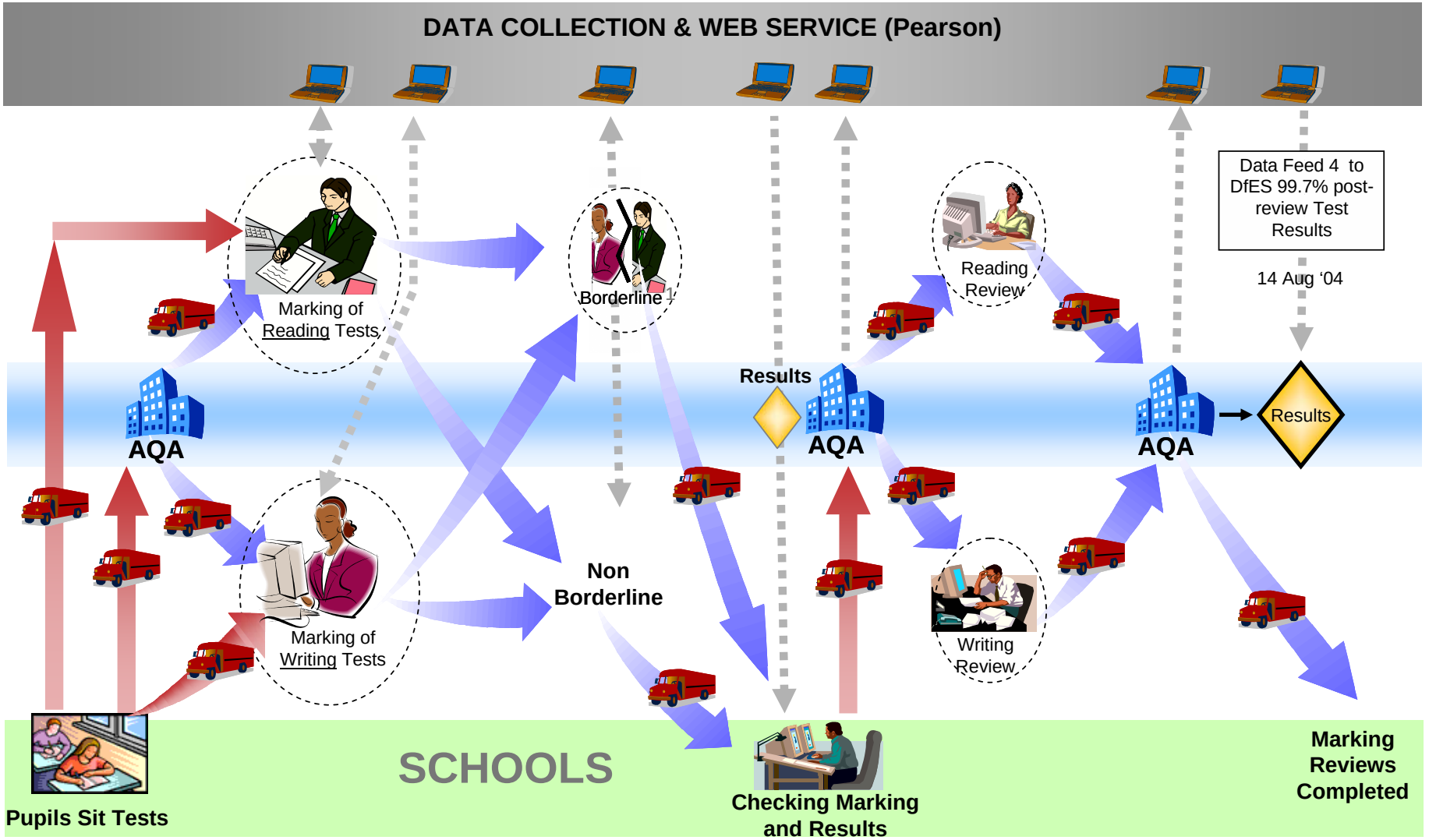
Recommendation 3: the team should be co-located and held accountable to develop and deliver a robust detailed process in a manner that allows it to become an ongoing operational process for subsequent years.

The detailed process developed by the team must be tightly scheduled and modelled with a full range of historical variables in order to reduce risk. Given the high stakes nature of the 2004–2005 test delivery cycle it is recommended that the team reports monthly through the QCA Chief Executive to the QCA Board using an agreed status report.

There are many recognised project management tools available. It is recommended that one is chosen which allows the team to rigorously track progress in order that it can be demonstrated on a monthly basis to the QCA Board that test operations are proceeding to plan. It is likely that such a methodology will use a strong 'gateway' or 'milestone' structure, rigorously enforcing the project team's achievement of tasks to timescales.

Figure 3

KEY STAGE 3 ENGLISH – PLANNED MARKING AND REVIEW / BORDERLINE PROCESS 2004



Component Marking and Subject Level Borderlining

The review team found that there was widespread support for component marking which was introduced in 2003–2004. Early evaluation suggests it was a positive innovation addressing, to an extent, one of QCA's priorities: the need to improve the quality of marking.

The chart at Figure 3 reflects the planned process. There was no convincing evidence that the timescale for the marking in 2004 was inadequate overall, however it must be noted that Data Feed 4 to the DfES has not achieved 99.7% **post review** test result contractual completeness in recent years. The delays arose from the poor management of the associated logistical and coordination problems detailed in Annex A.

However, in 2003–2004 full subject borderlining, coupled with component marking, gave rise to significant logistical and data capture issues that resulted in late, unreliable and incomplete electronic results to schools and the return of unborderlined scripts to some schools.

It was clear from a senior NAA manager's evidence to the review team that it was not until Friday 16 July that it was understood the problem was more serious than previously expected. The manager's understanding, from as early as 5 July, was that 90% of scripts had been completed by this date, based on the reliance the manager had placed on management information provided by AQA and Pearson.

The review team's investigation confirms that, at this stage of the process, accurate and reliable management information was not available or provided by NAA's suppliers. Nevertheless, irrespective of the perceived extent of the problem, the review team believes that it was a mistake for the NAA to return any pupils' scripts to schools unborderlined. This should not happen again in the future.

Recommendation

Component marking should be continued as it reflects wider examining practice, offers the potential for greater reliability and has generally been welcomed.

Recommendation 4: **in order to make the end-to-end delivery of the key stage 3 English test more robust QCA and DfES should seriously consider implementing component borderlining.**

The review team believes that component borderlining will serve to reduce logistical complexity and, according to initial indications, further improve marking quality and reliability. In line with all process changes the proposals will require a thorough programme of modelling to reduce risk.

Annex A

Specific Errors and Issues in the 2003–2004 Key Stage 3 English Test Cycle

Markers and Markscheme

- Historic lack of markers for key stage 3 English.
- Some markers had been recruited beyond the 31 March 2004 cut-off date for marker recruitment. An extra training session for markers had to be organised by AQA at short notice, and some of the marker materials did not arrive in time for the training session. Some markers were therefore not familiar with the markscheme before the training session.
- Some marker materials were sent to the wrong addresses due to the use of a flawed AQA database.
- Some of the test packs sent out by QCA's distribution unit contained the incorrect markscheme requiring AQA to send out additional copies to markers before they could start their pre-training activities.
- Late sign-off by NAA of school and marker documentation led to AQA having to omit key details from these administrative materials.
- Software on the AQA marker database was flawed causing some scripts to be sent to markers at the wrong addresses, with consequent delays.

Risk Analysis and Management Information

- PricewaterhouseCoopers, appointed by NAA to process-model test operations initiatives for 2004, was appointed too late in the 2003–2004 test cycle.
- NAA did not provide PECS (Production Engineering Consultancy Services), the company subcontracted by Pearson for data entry processing, with capacity modelling software that had been provided by QCA in previous years.
- PricewaterhouseCoopers was given non-historical data by NAA to model the 2003–2004 key stage 3 English test process, particularly in respect of marker script completion rates. There was also a delay by NAA in supplying the information needed for the model.
- Inadequate management information processes were available for tracking progress on marking, borderlining and reviews, all key elements of the critical path of the process.

Procurement

- The National Curriculum Test Operations procurement exercise, taking place in summer 2004, meant that key NAA Test Operations staff were deployed to work on the 2005 test contract, as well as trying to resolve issues with the 2004 test process.

- There was a poor working relationship between the two external agencies, AQA and Pearson, particularly during and immediately after the procurement exercise.

- Anticipating the 2004 procurement exercise, AQA's decision to restructure on a national basis meant this additional change should have been factored into the test cycle and risk management.

E-Results

- The late decision by Pearson, agreed by NAA, to deliver results online via the qcaupdate website, as opposed to delivering results via the School to School (s2s) website led to poorly planned implementation.

- An NAA letter informed schools that 'results' would be published on the qcaupdate website from 7 July. The published data available to schools from 7 July were 'live', incomplete and subject to updating and therefore unreliable.

- There was no information on the qcaupdate website about the status of the data or warning that data could be incomplete and subject to revision.

- Wrong access codes and passwords for the qcaupdate website were issued to schools.

- The high volume of users trying to access the qcaupdate website led to a website failure that should have been anticipated by Pearson. More bandwidth had to be added.

Subject Level Borderlining

- A decision made by NAA late in the test cycle to mark at component level and borderline at subject level did not allow adequate time for risk assessment and effective implementation.

- Flawed Pearson borderlining software resulted in a significant number of pupils' scripts being borderlined that should not have been.

- Markers were unable to access the Pearson online borderlining database as the ID and passwords required to log-on had been sent out the wrong way around.

Other Issues

- Poor mechanisms to escalate issues between all parties.
- Poor communications with schools concerning borderlining, the return of scripts and e-results.
- Insufficient meetings of the Test Operations Board.
- Failure by NAA and QCA divisions to establish a crisis management team once it was clear there were serious test delivery problems.
- No regular, rigorous monitoring or accountability forum.
- No systematic evaluation of feedback from schools, QCA divisions, DfES, LEAs.
- Late changes by QCA to the specifications of aspects of the test delivery process were not worked through with suppliers before gaining agreement from the DfES.
- The Distribution, Data Collection and External Marking Agency helplines failed to properly handle the volume of calls received.

Annex B

Review Team Observations

The review team observed several areas for improvement that DfES, QCA, NAA and schools may wish to consider.

- **The School Exam Office**

Unlike other public exams, where materials and guidance are clearly addressed to the exam office and have a centre code on the outside of the packet so that the school post room can identify them, key stage materials and guidance are variously addressed to the Headteacher, the KS3 Coordinator, the Assessment Coordinator, heads of English, Maths and Science, and even the Head of Year 9.

Recommendation: all key stage materials and guidance are clearly labeled and sent to the exam office in secondary schools.

- **One Point of Contact for Schools**

Schools are asked to contact each different agency involved in delivering a key stage test for different, specific purposes. Schools do not need to know how many different agencies and sub-contractors there are. For example, it is time consuming and frustrating for schools to work out whether to contact NAA or Pearson over particular queries.

Recommendation: schools should have one point of contact.

- **Entering Pupils for National Curriculum Tests**

It is useful for state schools to use the Pupil Level Annual School Census (PLASC) data when entering pupils for the key stage tests. However, using this data does not accommodate independent schools.

Recommendation: the review team believes that NAA, as part of their modernisation programme, should investigate the feasibility of a whole-school test and exam entry system so that entries for key stage 3 tests are carried out in the same way as GCSE and GCE, and administered by a school's exam office.

A clear cut-off date for pupil entry applications also needs to be established, with further arrangements made for adding pupils to the entry list who subsequently change school.

- **Borderlining**

The review team could find no overwhelming logic why pupils' scripts that are within three marks *above*, as well as below, the whole subject level threshold are not borderlined.

Marking Quality Improvement

- **Marker Comparison**

The review team heard evidence suggesting that, on completion of the key stage 3 English test, and after schools had sorted their pupils' scripts into the reading and writing components of the test ready to send to the respective reading and writing markers, there could be a further, random, halving of pupils' scripts—perhaps based on pupils' surnames.

A writing marker, for instance, would therefore receive half their normal writing component script allocation from one particular school. That marker's total marking load for a particular test cycle would, however, remain unchanged in comparison to previous years.

This technique would have the benefit of allowing the Lead Chief Marker of the External Marking Agency to make a comparison of the average marks of the two halves of a school's writing component, for instance, which would be an effective tool in helping to gauge marking quality. It was further understood by the review team that this would be a better technique than relying on reviews upheld by the External Marking Agency, expressed as a percentage of the whole cohort.

- **Reviewing Schools' Results before Returning Scripts to Schools**

The review team felt that, in the longer term, it would be worth considering how the test system could be designed so that a more considered approach to the process of seeking reviews of pupils' scripts could be implemented. Although electronic results for schools were introduced, unsuccessfully, for the first time in 2003–2004 schools were also supposed to receive their pupils' scripts back by 13 July, before the end of the summer term, although only 77% of schools received their scripts back by this date.

At the moment the test system is designed around schools receiving their pupils' marks and National Curriculum levels from markers sending scripts directly back to schools before the end of the summer term.

This leaves very little time for the External Marking Agency to carry out further marking quality checks on pupils' scripts. In future, if historic National Curriculum level data for individual schools were made available to the External Marking Agency, it would give the agency another tool to identify, review and resolve any results that appeared anomalous based on historic trends *before* pupils' scripts are sent back to schools.

It was suggested that schools receive only their pupils' National Curriculum levels before the end of the summer term and receive their pupils' scripts separately perhaps after the summer term has finished. The review team considered that this could lead to a more considered appeals process.

In the longer term consideration should also be given to reviewing the comparison between the teacher assessment and the test outcome. Variations should be reviewed to check further the reliability of the test outcomes.

Recommendation: that further consideration be given to how a mechanism could be introduced to check schools' key stage 3 English test results against schools' historic National Curriculum level data, *before* pupils' scripts are returned. Pearson and NAA should give further consideration to any changes to the review process that may be required as a result.

- **Teacher Assessment and Review Requests**

On the basis that teacher assessments are submitted prior to receiving National Curriculum test results, the review team heard evidence suggesting that teacher assessment could usefully be another criterion incorporated into the process whereby schools request reviews of their pupils' scripts. Teachers requesting a review would enter the pupil's Teacher Assessment level on the necessary documentation before submitting the request.

Recommendation: in the longer term, consideration should be given to the use of teacher assessment as a criterion for schools to use when submitting review requests. Further consideration should also be given to whether there needs to be a cut-off date for review requests.

- **Key Stage 3 Science**

In 2003–2004 the review team noted that there were also some delays in delivering key stage 3 Science tests to some schools.

- **Regulating the National Assessment Agency**

The QCA Regulation & Standards Division should clearly define its regulatory role in relation to the NAA. It is essential that this QCA division remains outside of the co-located test delivery management structure.

- **Key Stage 3 English Markscheme**

The markscheme became more complex in 2003 for technical reasons associated with the test structure and, in part, as an attempt to reduce the subjectivity of English marking. In 2004 the markscheme was split as a result of component marking and in order to improve the manageability of the marking task. However, the nature of English as a National Curriculum subject means that, in an attempt to achieve consistent objective results, the marking structure has become highly atomised.

Annex C

Abbreviations

AQA	Assessment and Qualifications Alliance
COI	COI Communications
DCA	Data Collection Agency (Pearson)
DfES	Department for Education and Skills
EMA	External Marking Agency (AQA)
LEA	Local Education Authority
NAA	National Assessment Agency
PECS	Production Engineering Consultancy Services
PLASC	Pupil Level Annual School Census
QCA	Qualifications and Curriculum Authority