An Evaluation of External Marking Review Services During 2000

A report for the Qualifications and Curriculum Authority

FULL REPORT

Paul Newton and Chris Whetton

22 December 2000
Acknowledgements

The authors would like to acknowledge and thank members of the NFER Project Team: Angus McDonald (Assessment and Measurement); Debbie Banks and Bunmi Idowu (Data Entry); Simon Rutt and Peter Emery (Statistics); Christine Webster and John Hanson (Field Research Services); Mohinder Rattu (Secretariat); and Pat Ashby (External Consultant for the case study research).

We would also like to acknowledge and thank other members of NFER staff who contributed: Hannah Crawford and Helen Benzie (Assessment and Measurement); Mary Minnis and Humaira Aslam (Field Research Services); Andy Mackinson, Liz Gibson, Pat Bhullar and Jackie Hill (Secretariat); and Pauline Benefield (Library).

It would not have been possible to conduct the evaluation project without the willing assistance of the many professional ‘stake-holders’ who contributed evidence. We would like to thank the following, in particular, for the time and effort that they put into the project: the national curriculum Chief Markers and Lead Chief Markers; the External Marking Agency project managers; the Qualifications and Curriculum Authority external marking project managers; and representatives from the National Data Collection Agency and the Department for Education and Employment.

We also extend our thanks to representatives from the following awarding and regulatory bodies who provided important documentary evidence: the Assessment and Qualifications Alliance; Oxford, Cambridge and RSA Examinations; the Edexcel Foundation; the Welsh Joint Education Council; the Northern Ireland Council for Curriculum, Examinations and Assessment; and the Examinations Appeals Board.

Finally, we would like to thank the many schools that provided evidence relating to the evaluation.

Paul Newton and Chris Whetton
22 December 2000
Contents

1. Introduction to the evaluation of external marking review services during 2000 1
   1.1 The external marking review system 1
   1.2 The External Marking Agencies 2
   1.3 The evaluation 2

2. Procedures for processing KS2 and KS3 review requests in England during 2000 (Study 1) 5
   2.1 Introduction 5
   2.2 Methodology 5
      2.2.1 The participating agencies 5
      2.2.2 The procedural documentation sought 6
      2.2.3 The interviews 7
   2.3 Results 8
      2.3.1 Exploring procedures for the processing of review requests by the English EMAs 9
      2.3.2 Comparing procedures for the processing of review requests by the English EMAs with those employed by other agencies 26
   2.4 Summary 37
      2.4.1 Inconsistencies between the two key stages 38
      2.4.2 Inconsistencies between the three national curriculum test subjects 38
      2.4.3 Inconsistencies between the three English EMAs 39
      2.4.4 Differences between the EMAs and the examination boards 40

3. An analysis of data relating to the external marking review services (Study 2) 43
   3.1 Introduction 43
   3.2 Methodology 43
      3.2.1 Aggregate review request data 44
      3.2.2 Individual review request data 44
      3.2.3 Aggregate EuR request data 46
   3.3 Results 46
      3.3.1 Aggregate review request data 46
3.3.2 Individual review request data
3.3.3 Aggregate EuR request data

3.4 Discussion
3.4.1 Quality of review data
3.4.2 Overall impressions from the review data
3.4.3 Patterns that emerged from the review data

4. An investigation into schools’ views of the national curriculum review system (Study 3)

4.1 Introduction

4.2 Methodology
4.2.1 Sample selection
4.2.2 Questionnaire development
4.2.3 Questionnaire distribution

4.3 Results
4.3.1 Response rates
4.3.2 Reply-sample representativeness
4.3.3 Responses to the survey questions
4.3.4 Comparing responses between the External Marking Agencies

4.4 Discussion
4.4.1 What do schools think of the quality of initial marking?
4.4.2 How do schools check the marking of test scripts?
4.4.3 What motivates schools to request reviews?
4.4.4 What discourages schools from requesting reviews?
4.4.5 Who do schools involve in the decision to request a review?
4.4.6 Do schools understand the review system?
4.4.7 What do schools think of the review system?
4.4.8 What do schools think of the EMAs?

5. A ‘customer satisfaction’ perspective on national curriculum review services during 2000 (Study 4)

5.1 Introduction

5.2 Methodology
5.2.1 Questionnaire sample selection
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2.2</td>
<td>Questionnaire development</td>
<td>105</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Questionnaire distribution</td>
<td>105</td>
</tr>
<tr>
<td>5.2.4</td>
<td>The follow-up telephone interviews</td>
<td>106</td>
</tr>
<tr>
<td>5.3</td>
<td>Results</td>
<td>107</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Response rates</td>
<td>107</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Reply-sample representativeness</td>
<td>107</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Responses to the survey questions</td>
<td>108</td>
</tr>
<tr>
<td>5.3.4</td>
<td>The ten case study telephone interviews</td>
<td>133</td>
</tr>
<tr>
<td>5.4</td>
<td>Discussion</td>
<td>136</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Characteristics of schools that requested reviews</td>
<td>137</td>
</tr>
<tr>
<td>5.4.2</td>
<td>The process of requesting reviews</td>
<td>139</td>
</tr>
<tr>
<td>5.4.3</td>
<td>The service provided by the EMAs</td>
<td>142</td>
</tr>
<tr>
<td>5.4.4</td>
<td>Reactions to review outcomes</td>
<td>143</td>
</tr>
<tr>
<td>5.4.5</td>
<td>Impressions of the review system</td>
<td>144</td>
</tr>
<tr>
<td>6.</td>
<td>Professional perspectives on the processing of KS2 and KS3 review requests: strengths, weaknesses and scope for improvement (Study 5)</td>
<td>147</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>147</td>
</tr>
<tr>
<td>6.2</td>
<td>Methodology</td>
<td>147</td>
</tr>
<tr>
<td>6.3</td>
<td>Results</td>
<td>149</td>
</tr>
<tr>
<td>6.3.1</td>
<td>The general functioning of the review system</td>
<td>149</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Procedures for informing schools of their right to a review</td>
<td>156</td>
</tr>
<tr>
<td>6.3.3</td>
<td>Procedures for the administration of review requests</td>
<td>158</td>
</tr>
<tr>
<td>6.3.4</td>
<td>Procedures for selecting, training and standardising review markers</td>
<td>162</td>
</tr>
<tr>
<td>6.3.5</td>
<td>Procedures for undertaking review marking</td>
<td>167</td>
</tr>
<tr>
<td>6.3.6</td>
<td>Procedures for reporting outcomes of reviews to schools and the NDCA</td>
<td>171</td>
</tr>
<tr>
<td>6.3.7</td>
<td>Procedures for dealing with complaints against, or concerns with, the processing of review requests</td>
<td>175</td>
</tr>
<tr>
<td>6.3.8</td>
<td>Procedures for auditing, monitoring, or evaluating the effectiveness of the review process and for determining ‘customer satisfaction’</td>
<td>177</td>
</tr>
<tr>
<td>6.3.9</td>
<td>Overall impressions of the system</td>
<td>178</td>
</tr>
</tbody>
</table>
7. An evaluation of external marking review services during 2000

7.1 Introduction

7.1.1 Did each of the EMAs run a satisfactory operation?

7.1.2 Did schools use the system in a satisfactory manner?

7.1.3 The evaluation process

7.2 A framework for evaluation

7.3 Interpretation and specification of ‘due process’, documentation of principles and of practices

7.3.1 The rationale of the review system

7.3.2 The formal documentation of procedures and practices

7.4 Transparency, public understanding and user-friendliness

7.4.1 Transparency and public understanding

7.4.2 User-friendliness

7.5 Openness, independence and trustworthiness

7.5.1 Openness

7.5.2 Independence

7.5.3 Trustworthiness

7.6 Technical effectiveness of structures and procedures

7.6.1 Procedures for informing schools of their right to request a review

7.6.2 Procedures for the administration of review requests

7.6.3 Procedures for selecting, training and standardising review markers

7.6.4 Procedures for undertaking review marking

7.6.5 Procedures for reporting outcomes of reviews to schools and the NDCA

7.6.6 Procedures for dealing with complaints against, or concerns with, the processing of review requests

7.6.7 Procedures for auditing, monitoring, or evaluating the effectiveness of the review process and for determining ‘customer satisfaction’
7.7 Between the present and the future
7.7.1 An alternative review model

8. References

| Appendix 2.1 | Documents received from the QCA, EMAs and DfEE relating to national curriculum test reviews |
| Appendix 3.1 | Adjustments made to the EMA databases prior to analysis |
| Appendix 3.2 | The study 2 results tables |
| Appendix 4.1 | The four Study 3 questionnaires |
| Appendix 4.2 | The Study 3 results tables |
| Appendix 5.1 | The seven Study 4 questionnaires |
| Appendix 5.2 | The ten follow-up case studies |
| Appendix 5.3 | The Study 4 results tables |
| Appendix 6.1 | The Generic Interview Schedule |
Section 1 Introduction to the evaluation of external marking review services during 2000

It is practically impossible to achieve perfect reliability of marking in large scale tests. Employing humans to mark scripts introduces numerous risks: they may forget to mark certain questions or may add scores up incorrectly; they may reach alternative interpretations of the mark scheme; or they might simply judge the quality of similar responses differently.

Although we cannot expect national curriculum tests to be marked with perfect reliability, we are entitled to expect that a respectable degree of consistency be achieved. To this end, the QCA, through the External Marking Agencies (EMAs), ensures that markers are thoroughly trained to apply the published mark schemes accurately. In addition, initial marking is monitored through a hierarchical system to ensure that markers have imbibed their training thoroughly. Despite these rigorous mechanisms, errors inevitably creep into the system. As the results are important for both pupils and schools, it is only fair that a mechanism should be in place to rectify such errors.

1.1 The external marking review system

The national curriculum tests in English, maths and science are taken by around 600,000 pupils at key stages 2 and 3, respectively. Each year, around 11,000 markers are appointed by the external marking agencies to mark these scripts. As the tests are administered in May, the marking schedule is very tight to ensure that scripts are returned to schools before they break up in July. Scripts are returned directly to schools by markers. For 2000, the deadline for returning marked scripts to schools was 7th July.

Schools are not required to check the marking of their pupils' scripts, although - particularly if there are obvious discrepancies between test and teacher assessment levels - it is advisable for them to do so. If - having checked pupils' scripts - a school is concerned that one or more pupil may have been awarded an incorrect level, it is at liberty to lodge a review request with the EMA responsible for marking its scripts. For 2000, the deadline for submitting reviews requests was 19th July.

Each of the External Marking Agencies offers a range of post-testing review services, to ensure that pupils who have been awarded incorrect levels can be awarded correct ones.

---

1 That is, unless computerised scoring is employed, which tends to require the kind of multiple choice format, or automated process, that many see as a threat to validity.
These are similar in purpose to the post-examination enquiry-upon-results services offered by the GCSE/A/AS examination boards. One of three types of review request may be made:

1. **R1 review.** A script check, for individual pupils, that determines whether clerical tasks (such as mark totalling) have been completed appropriately;

2. **R2 review.** A script check, for individual pupils, that determines whether the mark scheme has been applied appropriately;

3. **R3 review.** A script-batch check, for a group of pupils within a school, that determines whether the mark scheme has been applied appropriately. (This is only permitted for KS3 English.)

The R1 reviews are undertaken internally by EMA staff and most can be despatched to schools before the end of the summer term. The R2 reviews are undertaken by senior markers, selected from amongst the highest echelons of the marking hierarchy. They remark the scripts and send them back to the EMA. The EMAs typically retain these remarked scripts and return them at the beginning of the autumn term. The published date by which reviews should be completed was the 22nd September.

1.2 The External Marking Agencies

During 2000, two EMAs had responsibility for marking the national curriculum tests at key stages 2 and 3: the Assessment and Qualifications Alliance (AQA) and the Edexcel Foundation (Edexcel). In practice, the AQA had two largely independent sites, responsible for the north and south of England, respectively. These corresponded to the operations run in previous years by the Northern Examinations and Assessment Board (NEAB) and the Southern Examining Group (SEG), respectively. Because these operations had only recently merged and were effectively still running separately, the present report treats them separately and tends to refer to three EMAs: AQA-N (AQA north, based in Newcastle); AQA-S (AQA south, based in Guildford); and Edexcel (based in London).

1.3 The evaluation

If teachers are to retain confidence in the system it is essential that they remain satisfied with, and are prepared to accept the outcomes of, the review process. This means that the review system needs to be – and needs to be seen to be – highly effective. Procedures must be perceived to be **fair**, the External Marking Agencies must be trusted to operate
consistently and be seen to act impartially, and the process of review must be sufficiently transparent and comprehensible.

The QCA evaluation of KS3 external marking in 1997 touched upon the effectiveness of review procedures (Stobart, 1998). However, no evaluation project has focused specifically, or in any depth, upon review services. Following the external marking of tests in 1999, the Secretary of State for Education and Employment indicated that he wished to see an evaluation of review procedures undertaken to provide assurance of their fitness for purpose and to make any recommendations for change. In June of 2000, QCA invited tenders for a contract to conduct such a project. As specified in the Project Initiation Document, the aim of the project was to:

- examine the existing reviews procedures and to compare them with reviews and appeals services available for other tests and qualifications;

- provide assurance to QCA, the DfEE and schools that the reviews procedures are fair and operated consistently or to recommend changes which will improve them;

- contribute to the current consideration by QCA of its appeals procedures for all aspects of its operation of qualifications and tests.

More specifically, the objective of the project was to:

- provide a comprehensive survey of data relating to the reviews procedures and other related qualifications;

- provide an independent assessment of the operation of the reviews procedures from the perspectives of both schools and EMAs;

- provide an independent assessment of the comparability of the services offered by each EMA and the degree of consistency achieved, both in operational terms and the quality of review re-marking;

- assess how the revised role of the Lead Chief Marker for each subject has affected the operation of the reviews procedures;

- provide a report which highlights both strengths and weaknesses in the current procedures and makes recommendations for improvement.

It was clear that all of the above requirements could not have been achieved within the scope of a single study. Indeed, five inter-linked studies were proposed.
Study 1 was intended to explore marking review/enquiry/appeal procedures as they occurred for different UK assessment types. This was to involve predominantly desk-based research, comparing and evaluating documentation provided by test/examination agencies/authorities with responsibility for marking national curriculum tests, GCSEs and A/AS.

Study 2 was to focus upon data from the review/enquiry/appeal procedures for these agencies/authorities. Again, this was to involve predominantly desk-based research using information gathered from the QCA and from the test/examination agencies/authorities involved in Study 1. Where possible, the prevalence of review requests from different quarters in different contexts was to be explored, as was the prevalence of different review outcomes. These analyses would help to indicate whether requests, or specific outcomes, were more prevalent in certain circumstances.

Study 3 was designed to explore the reasons that schools had for requesting, or not requesting, reviews. It was assumed that these would not necessarily reduce to having questionably marked scripts. A survey methodology, with a random sample of schools, would explore a variety of issues, including: their grasp of the review system; their understanding of documents produced by the EMAs; the importance that they placed in maximising results; the amount of time available to them at the end of term; etc.. A better understanding of these issues would help to indicate whether the system was functioning fairly and transparently.

Study 4 was intended to focus more specifically upon only those schools that requested reviews during 2000. It would sample a large number teachers who requested either an R2 or an R3 review using a survey sent to schools early in the autumn term. This was to provide an evaluation of schools' satisfaction with the review procedures as they operated in 2000. Questions would touch upon satisfaction with: the processing of the request; the outcome; the turn-around-time; the deposit fee; etc.. An additional follow-up telephone survey of 10 schools was also planned, to explore some of the issues that arose in more detail.

Finally, Study 5 would enable the professionals involved in the review process to contribute their views on whether the system was functioning fairly, consistently, effectively and transparently. These views were to be elicited through one-to-one interviews and would explore both general issues and specific ones pertinent to the function of the representative within the system.

The NFER was awarded the contract to undertake this project on 5th July 2000 and began work immediately.
Section 2 Procedures for processing KS2 and KS3 review requests in England during 2000 (Study 1)

2.1 Introduction

The overall objective of the project was to evaluate the effectiveness of procedures for the review of national curriculum test results in England during 2000. As the first step towards this objective, Study 1 was intended as a preliminary investigation into how key stage 2 and 3 review requests were processed, both to frame the overall evaluative analysis and to identify issues for later studies. The first goal of Study 1 was exploratory – to identify the crucial documents, structures, processes and practices. The second was evaluative – to consider whether the documents, structures, processes and practices were effective. And the third was comparative – to compare procedures within the three English EMAs with those applied in other countries and in other qualification systems.

While Study 1 was essential in defining the scope of the project, it was not intended to be definitive in the sense of providing an authoritative statement concerning the effectiveness of the system. The five studies are to be seen as a whole, rather than as free-standing entities. As such, it was decided not to present any of the evaluative conclusions arising from Study 1 until findings from the other studies had been presented. The presentation of results for Study 1 therefore revolves around, firstly, a description of procedures followed by the English EMAs and, secondly, a comparison with procedures followed by the other agencies investigated. The issue of procedural consistency became central within this approach.

2.2 Methodology

To achieve the exploratory, evaluative and comparative goals of Study 1, the most suitable methodology was deemed to be one that combined a desk-based analysis of procedural documentation from a range of participating agencies with site visits to each of the three English EMAs.

2.2.1 The participating agencies

The following test/examination agencies were invited to participate in Study 1:

The three EMAs for England (re review procedures)

AQA-N Assessment and Qualifications Alliance – North (Newcastle site, ex-NEAB)

AQA-S Assessment and Qualifications Alliance – South (Guildford site, ex-SEG)
Edexcel          Edexcel Foundation

The EMAs for Wales and Northern Ireland (re review procedures)

WJEC          Welsh Joint Education Committee

CCEA          Northern Ireland Council for the Curriculum, Examinations and Assessment

The GCSE/A/AS exam boards for England, Wales and Northern Ireland (re enquiry-upon-results and appeal procedures)

AQA          Assessment and Qualifications Alliance

Edexcel          Edexcel Foundation

OCR          Oxford, Cambridge and RSA Examinations

WJEC          Welsh Joint Education Committee

CCEA          Northern Ireland Council for the Curriculum, Examinations and Assessment

Other agencies for England (re review and appeal procedures)

FAB          Examinations Appeals Board

QCA          Qualifications and Curriculum Authority

2.2.2 The procedural documentation sought

Letters to each of these agencies explained the rationale of the project, the demands that it would make, and the procedural documentation that would be required. The latter was exemplified as follows:

• relevant procedure file sections;

• instructional booklets, or explanatory material, for schools;

• training materials, explanatory documents or instructional notes for markers;

• pro formas for completion by schools, markers, officers, etc.;

• relevant memoranda (e.g., highlighting significant changes);
• duties and terms of appointment of markers;

• actual examples of reports sent to schools (and copies of report templates/standard reports, if available);

• progress/monitoring reports from previous years (for internal use, for QCA, etc.);

• research/evaluation reports from previous years (internal or external research).

Clearly, a wide range of documentation was sought from each agency. This was done on the assumption that the more documentation was received the better able the NFER would be to present a fair evaluation of procedural effectiveness. The agencies were asked to collate and return this information within three weeks.

None of the agencies declined to participate although, with the exception of the English EMAs, most had some difficulty responding within the desired time. Indeed, by the original deadline for receipt of documentation, there were still agencies that had not replied with contact names. To some extent this was understandable. The document collation was to take place during July 2000, one of the most busy periods for any national test or examination agency. This is all the more true with respect to departments responsible for reviews, enquiry-upon-results or appeals. Unfortunately, the incompleteness of the returns did compromise the scope for comparative analysis somewhat, particularly the comparison with Wales and Northern Ireland.

With respect to the QCA and the three English EMAs, the collation of procedural documentation was an iterative process. Without knowing in advance what documentation would be available from each agency it was not possible to make specific requests. However, as particular documents came to light from certain quarters it became apparent that comparable documentation might also be available from others. Through this iterative process documents that had not already been provided were requested from each of the English EMAs. In a similar way, the existence of key QCA documentation became apparent and these were also requested. Unfortunately, time constraints meant that this iterative process was not possible for the other agencies.

A list of all documents received from the QCA and from the English EMAs, relating to the review of national curriculum test results in England, is presented in Appendix 2.1.

2.2.3 The interviews

As the period during which Study 1 was conducted is such an intense one for the agencies involved, it was proposed to limit the interviewing of key personnel to only those
agencies that were the central focus of the present study: AQA-N, AQA-S and Edexcel. Site visits to each of the three English EMAs were conducted during late July. These tended to last around a day, in total, and involved:

1. interviews with external marking project managers and managers responsible for reviews;

2. being shown the locations and processes involved in review request administration;

3. collating outstanding procedural documentation.

Additionally, site visits to AQA-N and AQA-S involved observing elements of the re-standarisation meetings for KS3 English review markers.

It was made clear that the principal intention of the initial interviews was to clarify procedures that were employed by each agency. Subsequent interviews (Study 5) would explore in more depth the strengths and weaknesses of these procedures.

2.3 Results

The general impression of the intended and actual functioning of the system for reviewing national curriculum test results in England was positive. Each of the three EMAs demonstrated a clear commitment to dealing with review requests in a manner that was responsive to the needs of individual schools and consistent with the requirements of the QCA. The general impression from Study 1 was that review procedures were effective, fair and operated consistently between the three English EMAs.

It would be implausible to expect there to be no room for improvement, however. Therefore, in the spirit of constructive criticism, areas where there were question marks over the effectiveness, fairness or consistency of operation of specific aspects of the system are highlighted in the following presentation of results. If the report, at times, appears overly negative, this is not the conclusion that should be drawn; this is simply a consequence of choosing to highlight areas where improvements might be made.

The presentation of results is divided into two main sections reflecting two of the three components of Study 1: exploration of the English EMA procedures; and comparison of the English EMA procedures with those of other agencies. The evaluation of procedures was postponed until evidence from other studies had been presented.
2.3.1 Exploring procedures for the processing of review requests by the English EMAs

National curriculum tests at key stages 2 and 3 are the direct responsibility of the QCA. Rather than simply regulating these qualifications – as for GCSE/A/AS – the QCA has clear ownership of them. However, lacking sufficient administrative and technical resources, the QCA contracts out the test development, external marking and data collection to a number of distinct agencies. The relationship between the EMAs and the QCA is, therefore, technically one of contractor and supplier. This gives the QCA the right – should the QCA decide to exercise it – to specify quite precisely the manner in which external marking (including marking review) is carried out.

2.3.1.1 QCA principles

The most obvious vehicle for communicating required procedures to the EMAs is the contract for external marking, of which the Tender Specification forms a key component. Three versions of this document were available for scrutiny; they related to: the 1996-1999 contract for England; the 2000-2002 contract for England; and the 2001-2002 contract for London and Eastern England. Only the first of these documents specified procedures for the conduct of marking review in any detail. The six-page Annex N to the 1996-1999 Tender Specification set out as its purpose:

"to outline the procedures that EMAs should follow in handling review requests, dealing with review requests that do not comply with instructions sent to schools, monitoring the volume of requests and dealing with associated issues." (point 1)

It discussed eight aspects of the review system:

1. submission of review requests by schools;
2. monitoring volume of review requests;
3. procedures for processing review requests;
4. monitoring review markers;
5. principles for dealing with R1 reviews;
6. principles for dealing with individual R2 reviews;
7. principles for dealing with group R2 review requests;
8. standard documentation to be prepared.
Annex N combined specific requirements with general guidelines. For example, on the one hand, it specifically required that only Team Leaders, Senior Markers and Chief Markers may be used to conduct R2 review requests (point 7). On the other, it stated more generally that EMAs may derive their own systems for logging review requests in and out, as long as they were capable of producing information at any date on the volume of outstanding requests and on the status of individual school requests (point 6).

Subsequent specifications did not contain the same level of detail. The 2001-2002 specification did have an annex devoted to the review process, Annex K. However, it did not specify procedures to be employed, it merely related key facts. The 2000-2002 specification – the ‘live’ specification – had no annex devoted to the review process, although a number of review issues surfaced in other annexes. Instead, this document proposed that:

“Each year QCA will issue detailed procedures guidance to EMAs on the arrangements for processing reviews, which will include:

- logging of review requests;
- collection of deposits;
- following-up of incomplete documentation;
- appointment, training and standardisation of review markers;
- marking principles;
- standard documentation;
- reporting review outcomes to NDC Agency;
- reporting review outcomes to schools.” (point 3.8.2)

For 2000, this guidance was not issued in the form of a single document. Instead, the QCA’s intentions were communicated primarily through minutes of meetings with EMA managers and through emails intended to remind managers of procedures, to emphasise the importance of procedures, or to clarify the interpretation of procedures.

The QCA clearly recognises the wealth of practical experience within the EMAs and seeks to ensure that procedural decisions are made in consultation with EMA managers. Two important consultation points should be highlighted in relation to the processing of reviews for 2000. The first concerned the development of the Schools’ Guide, of which
Section 11 relates specifically to the review process. As this document explains how
schools should make review requests, it was essential that any procedural changes for
2000 that would have affected this process were decided before the Schools' Guide was
produced. The second consultation point was the meeting of 24 March 2000 in which
procedures for 2000 were finalised. This was an opportunity for EMA managers to bring
issues to the attention of QCA, for QCA to discuss issues with the EMAs, and for final
procedural decisions to be made. The minutes of this meeting were (re-)circulated in an
e-mail from the QCA, dated 5 July 2000, with a covering message that emphasised certain
of the agreed procedural changes.

2.3.1.2 EMA practices

The following discussion highlights eight stages in the review system and, for each one,
explains how the EMAs put the QCA's principles into practice.

2.3.1.2.1 Procedures for informing schools of their right to a review

The principal vehicle for informing schools of their right to a review was the Schools' Guide. This was circulated to schools prior to their receipt of national curriculum test
packs. The 2000 Schools' Guide — prepared by AQA-S as part of its contractual obligation — succeeded the 1999 Instructions to Schools. It was a clear improvement upon
its predecessor, both in terms of form and content. It appeared to be well structured, well
written, well formatted and well illustrated. It explained both what a school must do to
request a review and what the EMA would do with that request. Each of the EMAs
produced their own version of the Schools' Guide, using the same QCA-approved
template. The only changes were to agency logos and contact points; otherwise, form and
content remained consistent between EMAs.

All three EMAs explained that there was little, if any, face-to-face contact with schools or
LEAs to explain the review system. AQA-S used to target general advice (which included
discussion of review procedures) at an LEA level, but this had fallen by the wayside by
2000. AQA-N occasionally operates an informal, reactive policy of targeting areas with
large numbers of review requests: using one 'aggrieved' school as a venue, local teachers
are invited to attend a meeting in which marking and review processes are explained. It
was noted that this kind of public-relations exercise is not costed in the contract with
QCA.

None of the three EMAs devoted web-site space to external marking. However, both the
QCA and the DfEE web-sites had references to review procedures within discussions of
external marking. The QCA site linked to electronic versions of the Schools' Guides,
while the DfEE site referred to (but did not link to) this QCA site.
Finally, all three EMAs ran an external marking Helpdesk through which all calls regarding review issues were channelled.

2.3.1.2.2 Procedures for the administration of review requests

In each EMA, responsibility for the day-to-day administration of the review process fell to (one of) the external marking deputy project manager(s). However, there were subtle differences in the administrative structures of each EMA. At AQA-S, for instance, there was also a division of responsibility amongst senior staff along subject lines. Issues relating to review requests for each of the three subjects were referred to the senior member of staff with responsibility for that subject. This included the final processing of outcome letters/reports. The manager with overall responsibility for the review process at AQA-S also had responsibility for English, the subject for which most reviews occur. At AQA-N, the manager in charge of the review process was also the Lead Chief Marker for KS2 English. However, while there were different named points of contact for review issues in different subjects at AQA-N, overall responsibility for all subjects was assumed by the one manager. At Edexcel, overall responsibility for all subjects was assumed by the one manager who was also the point of contact for all review issues.

The initial processing of review requests within each EMA was carried out by a small team of permanent and temporary staff who followed a similar pattern of stages. These stages included:

1. receipt-date-stamping each request;
2. sorting requests by key stage, subject and type;
3. checking the completeness of each review request;
4. completing a tracking-form for each request;
5. acknowledging R2 and R3 requests;
6. logging details of each request on a database;
7. filing paper details of each request.

Physically keeping track of review requests is not a trivial matter, particularly for the bulky R3 packages. AQA-N was particularly well equipped in this respect and was able to deal with virtually all of the initial processing of requests in one large hut. Shelves along one wall of this hut were devoted to the storage of R3 requests, with a single box for each school, arranged in DfEE number order. The one problem faced by AQA-N was...
the lack of computer terminals in the hut, which meant that request details had to be logged elsewhere.

Each EMA had designed its own forms for tracking review requests (see Appendix 2.1). This meant that all requests were accompanied through all stages of the review process by a sheet of paper on which key details were recorded (e.g., dates of receipt and of key completion stages, school name and DfEE number, name of person processing request at different stages, etc.).

In accordance with the minutes of the meeting of 24 March 2000, all EMAs sent some form of confirmation of receipt of R2 and R3 requests. AQA-N provided acknowledgement letters which included all of the relevant text required by point 7 of the minutes of the meeting. (AQA-N distributed exemplar acknowledgement letters for R3 requests to the other EMAs for reference.) AQA-S also sent out full acknowledgement letters, although they did not cover all of the issues noted in point 7. Edexcel sent out postcards which acknowledged the bare details of each request but that, again, did not cover all of the issues noted in point 7.

In relation to the logging of request details on computer, all EMAs appeared to have databases that recorded sufficient detail in a user-friendly manner. AQA had recently installed a central system which was managed at AQA-S. Although this installation had been delayed, with particular knock-on effects for the logging of requests at AQA-N, the system appeared to be functioning at both sites. The main problem for AQA-N was a significant delay between review processing and database logging: R1 and R2 requests were only logged after the clerical check (i.e., after the review process had been completed). However, as full paper records accompanied each request, this did not appear to be causing significant administrative problems. Both Edexcel and AQA-S were logging requests far sooner and R2s were logged before being sent to markers. While all EMAs were technically equipped to record details of review turn-around times, there was no intention to analyse such data for 2000 and database logging procedures tended not to be set-up so as to specify key dates with precision.

The filing of paper-based records for each request was handled in a similar way by each EMA, with separate lever-files for each review type/subject/key stage combination. Deposit cheques, on the other hand, were filed differently. AQA-N filed cheques in a safe in the deputy manager’s office; AQA-S sent cheques to their accounts department with a copy of the letter of receipt. Edexcel filed cheques in a ‘Cheque File’ with a copy of the tracking sheet.
One of the issues that the QCA was keen to tighten for 2000 was the processing of review requests that did not fully comply with instructions given in the *Schools' Guide*. One aspect of this was sticking to the submission deadline of 19 July (or 26 July for specified exceptions). AQA-N marked a single ring around the date stamp for requests that were late by less than a week and a double ring for requests that were more than a week late. AQA-S recorded late receipt with a blue sticker. When asked their policy on dealing with late requests all three of the EMAs responded with mixed sentiments. While it was generally explained that late requests were not processed unless good reasons were given, there was also an underlying suggestion that certain 'borderline' late requests were processed. AQA-N, for example, explained that late requests for clerical checks were generally dealt with as these were so clearly the fault of the EMA.

All three EMAs had procedures for dealing with requests that were incomplete. AQA-S and AQA-N specified missing items, in acknowledgement letters, that would need to be provided within 10 days (AQA-N) or 7 days (AQA-S). To speed things up, AQA-S requested that missing signatures be returned by fax. Edexcel used a fax transmission sheet to specify all missing items. Once again, all EMAs gave the impression of being somewhat flexible on the return of missing items, preferring to begin processing the review requests immediately and to worry about the receipt of missing items later. Unfortunately, for AQA-S and AQA-N, this flexibility led to a certain degree of confusion when provisional figures for the number of reviews completed were requested by QCA in early September: it was initially unclear how to classify the small minority of reviews that had been completed by the EMAs but for which items such as deposit cheques were still outstanding.

One of the hardest issues that EMAs must arbitrate is the question of whether schools have given sufficient explanation for their R2 or R3 request. The *Schools' Guide* requires schools to "explain how it is considered that the original marker has not applied the mark schemes correctly" (for an R2 review). This is somewhat ambiguous and gives little indication of the level of detail required. Do these instructions explicitly proscribe a general explanation along the lines of "this child's work is clearly of a far higher standard than her marks for many questions would indicate"? Perhaps not. On the other hand, it would be hard to defend an explanation that stated simply "this child's script is clearly in need of a re-mark". In addition to the purely conceptual problem of how to decide whether an explanation is sufficient, EMAs also have to consider time implications. AQA-S explained that it was simply not feasible to go through every explanation to determine its sufficiency. Edexcel similarly tended to accept poorly expressed requests. AQA-N was, perhaps, somewhat more strict, typically requesting more explicit evidence.
for generally-worded requests, noting that the scripts would require full re-marking otherwise.

One further procedure for the administration of review requests is worth highlighting: the clerical check. In the minutes of the meeting of 24 March it was agreed that all scripts sent in must have, at least, a clerical check carried out on them. This, of course, is the essence of the R1 review. The procedure for clerical checking was most rigorous at AQA-S, where each R1 review was checked once before being passed to a second member of staff for a second check. Each check was recorded on the tracking form with the names of both members of staff involved. At Edexcel and AQA-N, scripts were checked only once, although at AQA-N there was also a sample check at data input stage. Each of the EMAs followed the same procedure for clerical checking R2 and R3 requests as they had followed for R1s. The clerical checks were conducted internally after scripts had been returned by the review markers.

The minutes of the meeting of 24 March were slightly ambiguous in the requirement that 'all' scripts should be clerically checked. AQA-N and Edexcel conducted full clerical checks on all scripts that were sent in to the EMA including, for example, even those scripts that were sent in as part of an R3 group for which a full re-mark was declined. AQA-S, on the other hand, clerically checked only those scripts that were actually subject to review (unless, in the rejected R3 scenario, there had been clerical concerns with the sample scripts).

Finally, although none of the EMAs had formal procedure files devoted to the step-by-step documentation of review procedures, each of them had developed in-house desk-instructions and or briefing notes for staff. The largest selection of instructions and notes was provided by AQA-N, which included, for example: written guidance for Helpdesk staff; written guidance on all stages of processing for all types of review request; and written statements of quality assurance principles (see Appendix 2.1).

Likewise, although the EMAs appeared to have no formal policy on recording departures from standard practice, there was evidence of good practice. AQA-N, for example, completed an 'RV Letter Log' which contained details of all errors made by schools when requesting their review. Edexcel similarly kept a log of which schools had sent in requests without mark sheets or without cheques. While AQA-S did not appear to have a central record of these departures from intended practice, details for each school were recorded on individual acknowledgement letters (copies of which were stored for reference purposes).
2.3.1.2.3 Procedures for selecting, training and standardising review markers

In accordance with decisions presented in the minutes of the meeting of 24 March, each of the EMAs selected only grade A supervising markers. AQA-N and AQA-S also stressed that markers should be recommended by the Chief Marker. AQA-S further specified that markers should not have any review requests against them. The number of markers appointed by each EMA for review purposes is presented below.

<table>
<thead>
<tr>
<th>Key stage</th>
<th>subject</th>
<th>Number of review markers appointed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Edexcel</td>
</tr>
<tr>
<td>2</td>
<td>English</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>maths</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>science</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>English</td>
<td>27 (+)</td>
</tr>
<tr>
<td></td>
<td>maths</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>science</td>
<td>2</td>
</tr>
</tbody>
</table>

Procedures for the training and re-standardisation of markers were not clearly defined for 2000. Even the minutes of the meeting of 24 March were ambiguous in this respect: they noted that each EMA had its own approach and reference was made to a document, prepared by the QCA on 9 March 2000, that discussed Lead Chief Marker strategies (but that actually had little to say about review procedures).

With the exception of English, it appeared to be assumed that training was either not necessary or that the provision of written training documentation was sufficient. AQA-S presented a number of exemplar documents to the meeting of 24 March, one of which represented their Instructions to Markers Conducting Re-marking. Edexcel and AQA-N reproduced this booklet for 2000, with only minor changes, as a general document applicable across key stages and subjects. In addition, AQA-N produced a further instruction booklet specifically for KS2 and KS3 English.

AQA-S reworked ideas in their exemplar booklet to accommodate changes for 2000 and produced a different version for each key stage/subject combination. This seemed to be the most appropriate approach, particularly as the original exemplar booklet confused
issues of procedure for group reviews between sections for KS2 English and KS3 English. AQA-S instructions for KS2 English omitted reference to group reviews.

The instructions to markers generally conveyed the subtle nature of the review process. It was made explicit that the function of the review was to "confirm the accuracy of the marking where this is appropriate", that "there is no initial assumption ... that the marker’s original marking is inaccurate" and that markers should not "search for extra marks or give the candidate the benefit of any doubt which would not have been appropriate at the original marking stage". The AQA-N instructions for markers of KS2 and KS3 English were even more explicit: “In particular, if your assessment of a question differs from that of the original marker you should consider very carefully whether this represents an error in original marking or an acceptable difference in marking.”

For mathematics, it appeared that the instructional booklet constituted all the training that markers in each EMA received. The average number of review markers for mathematics was two – typically the Chief and Deputy Chief Marker – which seemed to be taken as implicit justification for the lack of training and, as we will see below, for the lack of re-standardisation.

Standardisation refers to the process by which marking agencies ensure that all markers interpret and apply a mark scheme in the same way before they make progress on their allocation of live scripts. The standardisation of markers for national curriculum tests is a formal process which is governed by QCA-approved procedures (as laid down in handbooks for markers and supervising markers). Each marker must attend a co-ordination meeting (where mark scheme application is explained) and must have a ‘standardisation sample’ of scripts re-marked by a more senior marker. Re-standardisation, on the other hand, tends to be a less formal process which is intended to ensure that review markers are fully re-appraised of an initial mark scheme, having not marked live scripts for a number of weeks or months.

To some extent, the concept of re-standardisation had merged with that of training for review purposes. There was certainly no formal re-standardisation of markers in the sense required for original marking. Indeed, even when re-standardisation scripts were produced – for KS2 and KS3 English – it was not always clear how these were to be used or what action would be taken in relation to them. Re-standardisation scripts were prepared for neither of the remaining subjects at either key stage. Once again, it seemed to be assumed that the most senior markers would not need re-standardising. This was particularly assumed to be the case when, for AQA-S, the re-marking was to be conducted in one room on one (or more) day(s) with ample opportunity for discussion.
Key stage 2 English came closest to the formal standardisation of original marking. In a memo to EMA managers, the Lead Chief Marker for KS2 English explained that: “It is the intention of the LCM, in order to fulfil his duties with regard to reviews [Duty A10 from the Duties and Terms of Appointment], to re-standardise all review markers nationally by post”. Three re-standardisation scripts were supplied, to be marked remotely by each KS2 English marker; Sample Script Record Forms were to be returned by post. Unfortunately, it was not clear exactly what feedback would result from this process. Nor was it explicit exactly what should have happened had any significant problems occurred – a second sample? Although no feedback was given immediately, the Lead Chief Marker did return comments to the Chief Marker at each EMA recommending, where necessary, aspects of individual review markers’ work that might need to be monitored. At AQA-N, the re-standardisation process for KS2 English was reinforced through an initial consideration of two further scripts at the beginning of their review panel meeting.

Procedures for KS3 English were similar although less formal. Once again, three re-standardisation scripts were prepared and sent to EMA managers to be distributed to their markers. At this stage, procedures for AQA and Edexcel departed. AQA-N and AQA-S followed a similar regime. All KS3 markers were required to attend a review panel meeting (in Newcastle and Guildford, respectively) at which R2 requests and R3 samples were to be marked. These meetings began with discussion of the three re-standardisation scripts (which all participants had marked in advance). Hence, the process became more like a training exercise than a re-standardisation one. These discussions were led by the Chief and Deputy Chief Marker at each EMA and the Lead Chief Marker was present at both meetings to clarify any interpretative issues.

At Edexcel, there was no discussion of the re-standardisation scripts with the Lead Chief Marker, as the initial meeting was attended by only a small number of senior markers who had all been heavily involved in the Post-Completion Checking phase. However, the re-standardisation scripts were sent out to remaining markers. Sample Script Record Forms were then returned to the Chief and Deputy Chief Marker for checking (in much the same way that KS2 English scripts were returned to the Lead Chief Marker). In the letter, from Edexcel to KS3 review markers, which contained their re-standardisation scripts, it was explained that the quality of their marking was assumed to be acceptable; as such, they were instructed to commence review re-marking before receiving formal feedback on their marking of the re-standardisation scripts. This appeared necessary to ensure that all KS3 review requests would be completed to deadline, following an initial delay.
2.3.1.2.4 Procedures for undertaking review marking

In an email of 22 June 2000, AQA-N (the Lead Agency for English) asked project managers at AQA-S and Edexcel how they planned to conduct their reviews for English at KS2 and KS3. It explained that, at AQA-N, panel meetings were planned for English but that all re-marking of review requests for maths and science would be carried out remotely. The email explained that AQA-N reviews for maths and science would largely be conducted by the Chief and Deputy Chief Marker, with a 'cross-sample checking' procedure to ensure consistency of standards. Thus, having received completed reviews back from one maths or science marker, AQA-N would forward a sample of this work to another marker requesting that they: check the review outcome decision; complete the necessary columns of a sample check form for any contentious marks; write brief comments explaining their reasons. If this process revealed differences of opinion, the initial marker would be sent details of the sample checking marker’s comments and asked to telephone the sample checking marker to negotiate an appropriate decision. Where consensus could not be reached, the word of the Lead Chief Marker would be final.

The AQA-N panel meetings for KS2 English were held over two days, 20 and 21 July; for KS3 English they were held over three days, 26, 27 and 28 July. The intention of these meetings was to complete as many of the requests as possible with the Chief, Deputy and Lead Chief Marker in attendance to resolve any queries and to confirm review outcomes where necessary. For the KS3 panel, all R2 reviews were completed first, followed by the R3 samples. At both key stages, requests that were not completed during the meeting were completed remotely and returned by post (clarifying any queries with the Chief or Deputy Chief Marker by telephone). The work of markers during the panel meetings was informally sample checked by the Chief, Deputy Chief or Lead Chief Marker. Reviews completed remotely after the panel meetings were sample checked by the Chief Marker.

At AQA-S, review requests for all subjects were processed by teams of markers working centrally in panel meetings at Guildford. Once again, where there were too many requests to process during each of the panel days, markers processed remaining requests at home. There were no formal sample checking procedures during any of the panel meetings (on the assumption that the Chief and Deputy Chief Markers were on hand to discuss any problematic decisions) nor was there any sampling of reviews completed after these meetings. However, there was some informal sampling of re-marking during the KS3 English panel meetings. Moreover, the work of any markers not in attendance at the panel meetings was sample checked.

At Edexcel, it was initially intended to double-mark all KS3 R3 English samples during panel meetings in London. Unfortunately, this proved too time consuming and remaining
samples were reviewed remotely. In the end, all review re-marking for all subjects was
carried out remotely. While all review markers were encouraged to discuss any problems
by telephone with the Chief or Deputy Chief Marker, there were no formal sample
checking procedures. However, the KS3 English Chief and Deputy Chief Markers did
meet in London on 4 August to discuss and confirm R3 sample decisions on the basis of
markers' outcome reports and Sample Script Record Forms. (It was noted that many of
the Edexcel review markers had been involved in the Post-Completion Checking process
and were, therefore, likely not to require further re-standardisation.)

It is worth commenting specifically upon procedures for reaching decisions on KS3
English R3 reviews, based on the sample of 15 scripts. At AQA-N and Edexcel, markers
received the training document and exemplar Sample Script Record Forms provided by
the Lead Chief Marker. They were encouraged to make recommendations to the agency
concerning the best course of action (consulting with the Chief or Deputy Chief Marker in
cases where the decision was not clear-cut). These decisions would be reviewed and
confirmed by the Chief and Deputy Chief Marker (at AQA-N, in collaboration with the
EMA manager in charge of reviews). AQA-S did not provide markers with the training
documents and instead asked markers to bring their completed samples to the Chief or
Deputy Chief Marker, during the panel meeting, for discussion before any
recommendation was made. To enhance the consistency of R3 sample decisions, the
decisions at AQA-S were made by the Chief or Deputy Chief Marker in collaboration
with the EMA manager in charge of reviews.

Finally, in order both to speed up and to enhance the consistency of review marking for
KS2 science - as well as to enhance the consistency of reports to schools - a set of
standard responses for typical queries was produced by the Lead Chief Marker and
circulated to all EMAs by AQA-S (the Lead EMA for science).

Two major grey areas in the processing of review requests were identified. The first
concerned procedures for undertaking R3 requests in subjects other than key stage 3
English. The second concerned procedures following the rejection of a request for an R3
review.

In an email of 5 July, QCA reminded EMA project managers that:

"You should not make special offers to schools to do something for them outside
these arrangements [i.e., R3 for KS3 English only]. This does not mean that you
are not expected to exercise your responsibilities for high quality marking
however. For example, if having undertaken a number of R2 reviews for a school,
you/the chief marker believes there is something generally wayward with the
marking of a school’s scripts, you should not flinch from calling all those scripts in for a complete re-mark. This is not an R3, but a quality intervention initiated by you.”

In an email of 6 July from QCA, it was stressed that:

“There is no facility for schools to request a review for whole groups of pupils in any other but KS3 English. They should be advised to identify individual pupils and submit these for R2 review.”

There appeared to be no clear and consistent interpretation of this slightly ambiguous guidance. AQA-S appeared to be the most forthright in rejecting group reviews for subjects other than KS3 English. AQA-N and Edexcel appeared to be slightly less strict. AQA-N formally acknowledged the potential for full re-marking in other subjects in a letter to review markers:

“In the event of [group reviews] being requested for other subjects the request will be rigorously examined but it could be that you may be asked to carry out a full re-marking of a set of school scripts.”

In accordance with QCA guidance, AQA-N required that any school seeking a full re-mark in subjects other than KS3 English should submit its request using the appropriate R2 forms for individual pupils (also specifying a sample of 15). Edexcel appeared to be slightly less formal in its approach, not necessarily requiring schools to complete a large number of R2 forms for individual pupils.

The second grey area concerned procedure following the rejection of a request for an R3 review; in particular, how to deal with individual pupils from the R3 sample of 15 whose re-mark marks warranted the award of a higher level. Edexcel took the approach of formally and automatically amending the levels for these pupils and notifying schools that this had been done. Where this occurred, schools were returned their deposit cheque in full even though their R3 request was technically refused. Assuming that schools’ samples would contain the worst 15 scripts, Edexcel did not explicitly invite schools to submit any further concerns as R2 requests.

AQA-S and AQA-N chose not to amend pupils’ results formally and automatically. Instead, they returned all scripts to schools with the suggestion that further individual R2s might be requested (but without specifying which pupils’ levels would be likely to change). To speed up the subsequent processing, they retained the school’s deposit cheque pending the submission of further R2 requests.
2.3.1.2.5 Procedures for reporting outcomes of reviews to schools and the NDCA

The email of 5 July, from QCA to the EMA managers, specified that:

"outcomes reported to schools should be professional and written in consultation with the chief marker."

In the minutes of the meeting of 24 March, which were attached to this email, it was actually noted that the outcome reports should be written in consultation with the Lead Chief Marker. Although the QCA guidance did not seem to be very explicit in clarifying the role of the Chief Marker or the Lead Chief Marker, Action Point 7 from the minutes specified:

"LCMs informed that in partnership with their LEMA, they should develop the quality of reports and responses provided to schools."

This concurred with A11 of the Duties and Terms of Appointment of Lead Chief Markers:

"lead chief markers will be responsible for preparing standard responses for use in preparing reports to schools on the outcome of reviews and for working with the lead EMA to circulate these responses to all EMAs."

With the exception of standard responses (re individual questions) provided for KS2 science, there was little evidence that Lead Chief Markers in any of the subjects prepared and circulated standard review outcome letters or reports. The only evidence of generally circulated documentation came from the meeting of 24 March which included exemplar outcome reports from 1999 in a number of subjects.

Nor was there evidence that Chief Markers in any of the EMAs had a significant input into the preparation of review outcome letters or reports for schools (other than in their capacity as individual review markers). Instead, these tended to be (re)produced by EMA managers from reports prepared by individual review markers. The general approach taken across EMAs was to report successful changes back to each school with little elaboration, but to explain review rejections in more detail, focusing particularly upon any issues that may have been raised in the school’s initial cover letter.

It was not possible to conduct an in-depth analysis of the nature and quality of outcome letters and reports produced by the EMAs. Apart from the fact that this would have required a separate study in its own right, few outcome letters and reports had actually been prepared for 2000 while the research for Study 1 was being undertaken. The following comments reflect the general approach taken by each of the EMAs.
At AQA-N, the majority of letters and reports were compiled by the manager responsible for the review process (who was also the Lead Chief Marker for KS2 English). However, reports for KS3 English were compiled by the Lead Chief Marker for KS3 English who was based at Newcastle for two days per week. As for all three EMAs, these reports were based upon comments in individual markers’ review outcome reports which had been word processed where necessary. It was emphasised that all outcome letters and reports were individually prepared and specifically related to comments raised in schools’ covering letters. Moreover, a greater amount of time and effort was put into letters and reports for schools that had put more time and effort into their original request. The process of reporting to schools was a time consuming one for the EMA manager; he described it as his ‘August task’. Although outcome letters and reports were compiled by the manager responsible for reviews at AQA-N – the external marking deputy project manager – they were signed on behalf of the project manager. Where reviews were conducted after the end of the summer term, outcome letters and reports were not despatched until the autumn term.

At Edexcel, outcome letters and reports for all subjects were compiled by the manager responsible for the review process, in a similar manner as discussed for AQA-N. There appeared to be slightly more of an emphasis upon the production of standardised reports at Edexcel, to cut down the amount of time taken in preparing responses. R2 review markers’ reports were generally returned directly to schools, unless inappropriate comments were detected (in which case a new comment sheet would be produced minus the offending comments). Where R3 reviews were conducted after the end of the summer term, outcome letters were despatched immediately following review decisions, although full documentation was not despatched until the autumn term. All letters were signed by the manager responsible for the review process.

At AQA-S, the compilation of outcome letters and reports for each subject was carried out by the respective lead officer. This ensured that all letters and reports were written, and signed off, by a member of staff with direct subject knowledge. Lead officers at AQA-S faced a slightly more significant task in adapting review markers’ comments to address issues raised in schools’ covering letters. This was because review markers never saw schools’ covering letters and so could not address their own comments to them directly. The EMA chose to withhold covering letters as these often had a tendency to be negative and dismissive and risked putting markers in a bad frame of mind. AQA-S explained that reporting to schools was an extremely time consuming task. Finally, where reviews were conducted after the end of the summer term, outcome letters and reports were not despatched until the autumn term.
All EMAs fed data on level changes to the NDCA via an automated electronic data transmission system. This appeared to function effectively.

2.3.1.2.6 Procedures for dealing with complaints against, or concerns with, the processing of review requests

For 2000, there was no formal mechanism through which a school could appeal against the outcome of its review (as long as EMA procedures had been followed appropriately). Yet all EMAs noted that there had been problems in the past with schools persisting in their complaints. Action Point 6 of the minutes of the meeting of 24 March specified:

"QCA to agree a line where schools persist in review argument once its EMA has made a final decision."

In the minutes of the meeting it was agreed that, in such cases, as long as EMAs were satisfied they had followed all procedures, they should send a copy of the relevant file and a report to QCA.

AQA-N explained that schools which did persist in complaints would be sent a letter setting out the following four points: (i) the review process was complete; (ii) review procedures had been applied correctly; (iii) there was no formal appeal procedure; (iv) any further communication should be directed to the QCA.

A number of schools sent letters of complaint concerning external marking directly to individuals and agencies other than their EMAs. These included: the QCA, DfEE, MPs, the Secretary of State for Education; etc.. Indeed, these complainants often wrote to a number of individuals and agencies, which then required considerable discussion of individual cases between individuals and agencies and personal letters of response from all. In their responses to these complainants, the DfEE, in particular, took the opportunity to explain the rigour of marking and review procedures that might not have been apparent to schools otherwise.

2.3.1.2.7 Procedures for auditing, monitoring, or evaluating the effectiveness of the review process and for determining 'customer satisfaction'

For 2000, as in previous years, the principal mechanism for monitoring the work of each EMA was the Project Management Report. These presented the number of review requests received and processed by set dates and were provided to QCA, by each EMA, on a regular basis.

In addition, each year the EMAs, Lead Chief Markers and Chief Markers are required to provide reports on the functioning of the external marking system. For 1999, the reports
from Lead Chief Markers and Chief Markers tended not to discuss the review process in detail, if at all. In contrast, the four EMA reports for 1999 (SEG, NEAB, Edexcel and OCR) devoted between one and five pages specifically to review issues. Interestingly, however, there was a notable lack of consistency in the approaches taken. Edexcel, for example, concentrated primarily upon the prevalence of review requests, across subjects and key stages, considering possible explanations for apparent trends in terms of the reasons given by schools for making their requests (apparently based on a detailed empirical analysis). NEAB, on the other hand, concentrated mainly on presenting a factual account of procedures followed by the EMA in processing review requests during 1999. SEG’s report tended to be more interpretative, attempting to explain the prevalence of review requests, across subjects and key stages, in terms of a variety of school and EMA factors. Finally, OCR’s report was the briefest, simply presenting general comments on the processing of review requests.

Across EMAs, there was a general recognition that process evaluation is important. This was reinforced by bullet point 3 of the Instructions to Markers, which explained that one function of the review process is:

“To enable the Chief Marker and [the EMA] to identify matters that need to be addressed in test development; training in 2000 and in evaluating test development.”

Indeed, each of the EMAs explained how evaluation of the initial marking of scripts feeds back into the process of grading of markers and affects the decision as to whether they will be re-appointed. Review decisions and review markers’ comments were of particular importance in this respect and were consulted explicitly.

More generally, though, all three EMAs felt that the formal evaluation of review procedures was not dealt with as well as they would like it to be. The overriding concern was to ensure that the process was completed successfully in the first instance and subsequent reflection on issues arising tended to take second place. AQA-N referred to mechanisms such as orally reporting to the Senior Management Team on a weekly basis and working individually with staff to determine scope for improving practice. Yet all EMAs explained that there were few formal internal mechanisms for feeding past experiences into the development of the system in future years.

While none of the EMAs conducted ‘customer satisfaction’ surveys in relation to external marking, the QCA commissioned SMSR to conduct a survey of schools’ satisfaction with national curriculum assessment as a whole, part of which focused on the review process.
Of course, the present project is another example of evaluative research commissioned by QCA.

Finally, an evaluation report which specifically addressed the training and re-standardisation of KS3 English review markers in 2000 was produced by the Lead Chief Marker at the QCA’s behest.

2.3.2 Comparing procedures for the processing of review requests by the English EMAs with those employed by other agencies

A major intention of Study 1 was to consider the procedures that a variety of agencies, other than the English EMAs, employed for processing review requests. It was hoped that this might suggest better ways of operating the national curriculum review process in England. The agencies highlighted were: (a) the GCSE/A/AS examination boards in England, Wales and Northern Ireland; and (b) the key stage test marking agencies in Wales and Northern Ireland.

Unfortunately, while Wales and Northern Ireland agreed in principle to participate, information was not forthcoming from them in time to be integrated into the present report. (The exception was a selection of GCSE/A/AS forms and standard letters provided by WJEC.) As a consequence, the following presentation of findings relates only to procedures employed by the three English examination boards (AQA, OCR and Edexcel) for processing GCSE/A/AS enquiry-upon-results (EURs).

Within the remit of the project, it was not possible to explore EUR procedures to the same depth as national curriculum review procedures; in particular, less effort could be devoted to ensuring a full coverage of documentation from each board and no interviews were conducted to clarify procedures. Therefore, it is possible that subtle, though important, differences may not have been detected. However, it is hoped that the major points of divergence were identified.

2.3.2.1 Post-examination EURs

It is essential to preface this discussion by noting essential differences between the external marking systems employed by the examination boards and the test marking agencies. While the test marking agencies are actually divisions within two of the three English examination boards, the nature of national curriculum testing has led to principled differences in procedures operated between the systems. Most significantly, EURs evolved to deal with a system in which examination scripts were never seen by schools, while reviews evolved to deal with precisely the opposite scenario. To illustrate this, a school seeks an EUR when it is believed that a candidate would have performed at a
standard higher than that indicated by the grade awarded (an inference based on pupil ability). Conversely, a school seeks a review when it is believed that a candidate’s observed performance was at a higher standard than that indicated by the level awarded (an inference based on pupil performance).

For 2000, however, the context changed significantly for the examination boards with the requirement that photocopies of externally marked scripts should be made available to schools, on request, for all GCE subjects and for English and mathematics at GCSE. Notably, schools had to request these scripts – they were not provided automatically. However, in principle, examination scripts would have been available for scrutiny before schools decided to lodge their EURs. Indeed, one intention behind the access-to-scripts scheme was that – having seen exactly how pupils did perform – schools might be less inclined to lodge EURs. It remains to be determined the extent to which this proved true.

While the examination boards have been externally regulated by QCA (and predecessors) for a number of years, they have also regulated themselves through the mechanism of the Joint Council for General Qualifications. For 2000, the Joint Council issued guidance for centres (Joint Council for General Qualifications, 2000a) covering EUR services that would be made available by all examination boards for GCE A/AS, Special Papers, GCSE and GCSE (Short Course). These included the following:

Service 1  **Clerical Re-check.** A full clerical re-check including the provision of a statement of the marks by component for an individual candidate. *For components covered by the access-to-scripts scheme, centres may request this enquiry with return of script (and mark scheme).*

Service 2  **Re-mark.** A re-mark of externally-assessed components including a full clerical re-check and statement of marks by component for an individual candidate. *For components covered by the access-to-scripts scheme, centres may request this enquiry with return of script (and mark scheme).*

Service 3  **Re-mark and Report.** A re-mark of externally-assessed components including a full clerical re-check and statement of marks by component with a report on an individual candidate’s performance. *This service is only for components not covered by the access-to-scripts scheme.*

Service 4  **Re-mark and Report on a Group of Candidates.** A re-mark of externally-assessed components of a group of no fewer than 5 and no more than 15 candidates, including a full clerical re-check, statement of marks
by component and a general report on the work of the group. *This service is only for components not covered by the access-to-scripts scheme.*

**Service 5**  
**Re-moderation of Coursework with Feedback.** A re-moderation of the centre’s coursework marks, together with a brief report on the assessment of the coursework of the candidates, similar to that provided at the original moderation.

In addition to these generally available services, certain examination boards offered extra services. For example, in addition to specific schemes for Art and Art & Design, AQA also offered:

**Service 7**  
**Centre Review.** A re-mark of externally-assessed components, for centres where 10 or more pupils were entered for a subject and it was felt that the marking of all candidates was suspect. This involves script sampling and does not necessarily result in the re-marking of scripts for all candidates.

There are a number of ways in which these EUR services differ significantly from national curriculum review services. Many of the specific differences will be presented below under the headings used earlier to distinguish review stages. However, it will be important to begin by identifying a number of general issues.

### 2.3.2.2 General differences between EURs and reviews

Firstly, EUR re-mark services tend to be less defensive and to focus less on explanation and justification than national curriculum re-mark services. When grade changes occur, outcome letters for the basic re-mark service include only the most rudimentary of explanations why. When grade changes do not occur, no explanation is given. Where detailed reports are requested, these are intended more to highlight the general strengths and weaknesses of individual candidates, or groups of candidates, than to justify with precision the application of the mark scheme. Indeed, a report is effectively a substitute for access to a candidate’s script, which is why reports are no longer available for subjects involved in the access-to-scripts scheme. The lack of detailed explication of mark scheme application would seem to be a consequence of not having returned scripts in the past and, therefore, not having needed to justify specific applications. Yet, it persists into the present day when scripts are available for scrutiny.

Secondly, the language of EUR services often differs from that of review services, tending to be more authoritative. For example, centres are not asked to supply a ‘deposit’, but to pay a ‘fee’. The difference in emphasis empowers schools with respect to reviews, but empowers agencies with respect to EURs: the linguistic implication of a ‘deposit’ is that
it is likely to be returned, which effectively admits review marking error before it is
established. In contrast, the notion of a ‘fee’ assumes innocence until proven guilty.
Another subtle difference in terminology concerns the clerical check which, for EURs, is
correctly referred to as a ‘clerical re-check’. While perhaps more awkward, it conveys the
fact that all scripts were clerically checked even before final grades were reported. Yet
another difference concerns the description of those responsible for conducting the EUR
or review. While reviews are generally conducted by the ‘EMA’ or by ‘review markers’,
EURs are conducted by ‘senior examiners’. This more clearly conveys that those who
undertake the re-marking are those most qualified to do so.

Thirdly, although there are technically more EUR services than review services, there is
actually more scope for divergent practice within the review system. Essentially, at EUR,
scripts either receive a full clerical re-check or a full re-mark; there is generally no
intermediate stage of re-marking only certain questions. This makes the process less
ambiguous for all concerned. In addition, it means that there is no requirement upon
schools to explain how the examiner might have misapplied the mark scheme which
means less work for schools when completing their requests and no obligation upon
agencies to justify marks awarded.

Fourthly, there is no necessary provision for an R3-like review within the EUR system.
The Joint Council guidance would simply recommend the submission of multiple Service
2 re-mark requests for all pupils concerned. AQA, on the other hand, does recognise the
potential for requesting an enquiry that involves all candidates (Service 7). This begins
with sample re-marking in a similar way that an R3 review would. However, and
importantly, the sample is decided by the board rather than by the school.

Fifthly, the Joint Council guidance explicitly states that:

"Where an enquiry leads to a change of grade, normally only a higher grade will
be awarded. However, the awarding bodies reserve the right to award a lower
grade if the original grade had seriously misinterpreted the attainment of the
candidate in the subject." (Joint Council for General Qualifications, 2000a, point
5.4)

This contrasts directly with guidance, in the second paragraph of the Review Services
section of the Schools’ Guide, which warns:

"You should be aware that, depending on any errors found during the review, the
outcome could be that the marks and the pupil’s level go either up or down.”
(In fact, the Joint Council has decided that, from 2001, GCSE/A/AS candidates' grades will be lowered, should a clerical re-check or re-mark recommend this course of action.)

Sixthly, in the Joint Council guidance, there is explicit reference to action that would be taken by the examination board if an EUR highlighted a problem that might extend to other candidates (i.e., to some other candidates at the centre, to all other candidates at the centre, or to some or all other candidates at centres marked by a particular examiner):

“If the outcome of an enquiry is considered by the awarding body to bring into question the accuracy of the results for other candidates in that syllabus, the outcome of the original enquiry will be notified to the centre and the awarding body will extend the enquiry to include any other candidates who may have been affected. The centre(s) will be notified of the further investigation and the outcome.” (Joint Council for General Qualifications, 2000a, point 5.2)

Incidentally, OCR procedures state that, for modular subjects, the centre must give permission before the extension is commenced, as module marks can be taken up or down; the centre can also stipulate which candidates they wish to have as part of the extension.

There is no explicit provision for re-marking extension in the Schools' Guide; nor is it clear whether the QCA would expect EMAs to adopt this practice (particularly where scripts would need to be recalled from schools that had not formally requested any reviews).

2.3.2.3 Specific differences between EURs and reviews

Having discussed some of the general differences between EUR and review services, we will now consider some of the more specific contrasts. For ease of comparison, these are presented under the headings used previously to identify different stages of the review process.

2.3.2.3.1 Procedures for informing schools of their right to a review

For EURs, schools are directed to two principal written documents. The first is the Joint Council booklet described earlier. This presents information and guidance on all aspects of the EUR and appeal process and applies across all examination boards. The second is the booklet prepared by the examination board to which the EUR will be made. The board-specific booklets present similar information to the Joint Council booklet but also include application forms and more detailed information on how to prepare a request. They are sent into schools at the same time as examination results are returned. In
contrast to the *Schools’ Guide*, the Joint Council and board-specific booklets only contain information on post-examination services.

In addition to these booklets, the examination boards also discuss EUR and appeal services on their web-sites. Edexcel also distributes a poster which displays, in tabular form, the major post-examination services and key dates associated with them.

Within the board-specific booklets, the fees for individual EURs are presented. Instead of the uniform deposit required for national curriculum reviews, the boards charge different rates dependent on type of request, level of qualification, subject versus component request, etc. Indeed, different rates are charged between boards for the same services. The following figures, from AQA, illustrate the fees for considering all externally-assessed components for a subject (or all internally-assessed components re Service 5):

<table>
<thead>
<tr>
<th>Service</th>
<th>GCE A-level</th>
<th>GCSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service 1</td>
<td>£9.60 per candidate</td>
<td>£8.40 per candidate</td>
</tr>
<tr>
<td>Service 2</td>
<td>£39.00 per candidate</td>
<td>£23.20 per candidate</td>
</tr>
<tr>
<td>Service 3</td>
<td>N/A</td>
<td>£40.00 per candidate</td>
</tr>
<tr>
<td>Service 4</td>
<td>N/A</td>
<td>£125 for 5 candidates (+ £16.75 per additional candidate)</td>
</tr>
<tr>
<td>Service 5</td>
<td>£125.00 per centre</td>
<td>£112.50 per centre</td>
</tr>
<tr>
<td>Service 7</td>
<td>£310.00 per centre</td>
<td>£165.00 per centre</td>
</tr>
</tbody>
</table>

Clearly, EUR fees are significantly higher than review deposits (with the possible exception of Service 7, the R3-like service). However, it is not clear that the amount of work involved in processing each EUR is necessarily more than would be required for a comparable national curriculum review (reviews typically do not involve full re-marking, but they do often require a detailed outcome report). Note also that fees are charged for clerical re-checks. Finally, centres are not required to include fees with their EUR requests, as they are invoiced later. Edexcel and OCR invoice after completion of the service (but do not invoice when grade changes occur) while AQA invoices with the acknowledgement letter that is sent within seven days of receipt of request.

As mentioned earlier, forms for requesting EURs are presented at the end of the board-specific booklets. Typically, the boards provide only one generic form on which schools
record which service they require, for which pupil(s), in which subjects, at which level. (Edexcel and OCR request a separate copy for each subject while AQA requests a separate copy for each candidate.) This contrasts with the range of review request forms presented in the *Schools’ Guide*. The single form is the only method through which a school may request an EUR. If letters of concern about results are received, they “will be acknowledged and considered in the monitoring of centres’ responses to results” (Joint Council for General Qualifications, 2000a, point 2.1); but neither letters nor telephone calls will trigger an EUR.

Finally, schools were given the closing date of 20 September 2000 by which to have submitted their EUR requests – a month after receipt of examination results. Even given the time taken to request photocopied scripts from the boards, this was substantially longer than the time allowed for national curriculum review requests.

### 2.3.2.3.2 Procedures for the administration of review requests

EUR procedures for conducting clerical re-checks and re-marking were specified in the Joint Council booklet as follows:

**Clerical Re-check**

- a re-check that all parts of the scripts have been marked;

- a re-check that the marks on the scripts have been correctly totalled;

- a re-check that the total mark for each component has been correctly recorded in the computer system;

- a re-check that any examiner or moderation adjustments, if applicable, have been correctly applied;

- a re-check on the totalling of component marks;

- a re-check on the application of grade thresholds;

- a re-check as to whether any allowances, e.g., special consideration, have been applied;

- the provision of the mark for each externally-assessed component;

- a photocopy of the re-marked script if the component is involved in the access-to-scripts scheme (on request).
Re-mark

- a clerical re-check [consisting of all items presented above];

- an assessment of each externally-assessed component against the approved mark scheme for the component by a senior examiner of the original component who will confirm or amend the original marks.

2.3.2.3.3 Procedures for selecting, training and standardising review markers

Little information was provided by the boards on the selection of EUR examiners. AQA documents explained that examiners were allocated automatically from computer records and, for many subjects, involved only the Chief Examiner, Principal Examiners or Principal Moderator. Both AQA and OCR fees for EUR re-marking were generally 50% more than for prime marking, with additional fees for report writing.

In addition to the information booklet for centres, three instructional booklets were produced by the Joint Council for examiners (Joint Council for General Qualifications, 2000b, c and d). These contained instructions and guidance for examiners and moderators on re-marking, re-moderation and report writing. A further booklet was prepared for awarding body staff (Joint Council for General Qualifications, 2000e).

The booklet on re-marking (Joint Council for General Qualifications, 2000b) explained that, before undertaking any re-marking, each examiner must:

- re-familiarise him/herself with the paper and final version of the mark scheme;

- be familiar with the instructions and guidance included therein and, if requested, attend a post-results standardisation meeting (where more than one examiner would be re-marking scripts for a component).

It was further explained that examiners would normally have five working days, from the receipt of each script, to re-mark and return it. As for national curriculum reviews, examiners were instructed that they should not “search for extra marks or give the candidate the benefit of any doubt which would not have been appropriate at the original marking stage” (Joint Council for General Qualifications, 2000b, point 11). Similarly, staff were advised to convey to examiners that they should “not look for extra marks when marking under reduced pressure nor expect to find something wrong with the original marking simply because a result has been queried.” (Joint Council for General Qualifications, 2000e, point 20.4).
In addition to the Joint Council booklets, examiners received additional training material. AQA, for instance, produced a booklet containing distance learning materials for senior examiners involved in post-examination services (AQA, 2000). This covered similar ground to the Joint Council booklets (and included the first four of them as appendices). It ended with a section containing questions that examiners should have been able to answer having read the preceding sections. Although answers were provided, it was made clear that examiners should consult their Subject Officers if they had any lingering concerns.

As indicated above, examiners were not necessarily re-standardised, even when more than one examiner would be re-marking a component; the decision whether to re-standardise was at the discretion of the awarding body. No evidence on the actual prevalence of re-standardisation was provided.

2.3.2.3.4 Procedures for undertaking review marking

As for national curriculum clerical checks, EUR clerical re-checks were carried out internally by members of agency staff. Likewise, as for many review re-marks, many EUR re-marks were carried out remotely and returned by post. However, as for review re-marks at AQA-S, EUR re-marks in certain AQA GCSE/A/AS subjects were conducted in the Guildford office instead of being posted to examiners.

EUR re-marks, unlike review re-marks, were recorded using black ink to distinguish re-marking marks and annotations from those recorded by earlier examiners. Re-mark mark totals were written alongside the original mark totals and were initialled. Clear indication was required that each question and sub-question within a script had been re-marked. Although lower grades were generally not awarded, examiners were instructed that they should mark scripts down, as well as up, where appropriate. Examiners were not told how close candidates were to the grade boundaries of each component and would, therefore, not necessarily have known whether adding or subtracting marks would imply a change of grade.

In stark contrast to review re-markers, EUR examiners were encouraged to focus more attention upon candidates whose marks did change than upon those whose marks did not. In particular, EUR examiners were required to complete a Mark Change Justification Report (MCJR) for any candidate whose re-mark resulted in an overall mark change. These forms contained spaces in which to identify the questions that led to mark changes and spaces in which to record, from a list of five, the reasons for these changes (clerical error by original examiner/script checker, content mark(s) not rewarded/too generous, analysis mark(s) not rewarded/too generous, evaluation mark(s) not rewarded/too generous,
Examiners were also required to rate the quality of the original marking of these scripts (as good, satisfactory, poor or very poor) and were provided with a space on the MCJR in which they might record pertinent comments for the Subject Officer. Unlike reports written for national curriculum reviews, these EUR reports were intended solely for the purpose of the examination board.

One important issue to note concerned the prioritisation of EUR requests. In order to facilitate the processing of EURs, for candidates whose entrance to institutions of Further or Higher Education depended on their examination results, a priority service was provided by the boards. If such requests were lodged within a week of receipt of results, the boards promised to prioritise them and endeavoured to turn them around within 30 calendar days. Turn-around times for non-prioritised requests were 40 calendar days. Notably, all non-prioritised requests were dealt with on a first-come, first-served basis (in contrast to review procedures which prioritise clerical checks). This was made quite explicit for awarding body staff:

"With the exception of Priority Service 2 requests, applications for enquiry services should be dealt with strictly in the order in which they are received, with no attempt to prioritise for any reason, since it would not be appropriate for the awarding bodies to attempt to establish any criteria for dealing with individual requests out of sequence of their receipt." (Joint Council for General Qualifications, 2000e, point 10).

No indication was given, from documents received, that the re-marking of any EUR examiner was monitored in any way.

2.3.2.3.5 Procedures for reporting outcomes of reviews to schools and the NDCA

Outcomes of EURs were to be provided to schools within 40 calendar days of receipt of request (30 for priority requests). However, the boards made it clear that they did not guarantee these turn-around times. Outcome letters were despatched immediately following the completion of an EUR, while outcome reports (where requested) generally followed later.

The Joint Council guidance to staff (Joint Council for General Qualifications, 2000e, point 35) explained that outcome letters should, where possible, relate to the original enquiry and should be written with as personal a touch as is possible with a large scale operation. Furthermore, the guidance specified that a brief explanation for any grade change should be included, although this should not go beyond the reasons recorded on the MCJR (point 37).
As mentioned earlier, the Joint Council also provided guidance and instructions on the completion of reports to schools (Joint Council for General Qualifications, 2000d). It was made clear that they should be prepared in a form that could be sent directly to centres without further amendment or editing and that utmost care should be taken to ensure that they were accurate and fairly reflected candidates’ performances.

For candidates whose final grades changed as a result of re-marking, fees were returned in full. However, unlike review deposits, fees were only returned for candidates whose grades changed, not for additional candidates from a school whose work was submitted at the same time but whose grades did not change. Likewise, for Service 7 centre re-marks, AQA returned fees on a pro-rata basis.

2.3.2.3.6 Procedures for dealing with complaints against, or concerns with, the processing of review requests

In stark contrast to the national curriculum review system, the GCSE/A/AS system embraced no fewer than three formal mechanisms for appealing against EUR decisions: two of these were internal and one external.

The first internal procedure required an application from the head of centre within 14 calendar days of receipt of outcome. This would be considered by a senior member of awarding body staff who had had no previous involvement with the case. This officer would consider whether appropriate procedures had been followed and would be empowered to request a further re-mark if deemed appropriate.

If the head of centre was dissatisfied with the outcome of this stage, she would have another 14 calendar days within which to lodge a second internal appeal. The case would then be heard by a panel of three, one of whom would be independent of the board (i.e., would not have worked at that board within the past seven years). The panel would be empowered to decide whether procedures used were consistent with the regulatory authorities’ code of practice and whether they had been applied properly and fairly in arriving at judgements.

It is important to stress that the appeal process is not simply limited to procedures for marking scripts. It extends to procedures for: setting papers; marking and moderating; grade awarding; EURs and the appeal process itself (Joint Council for General Qualifications, 2000a, point 11.8).

If the appellant was still not satisfied, she would be able to lodge a further appeal with the independent Examinations Appeals Board (EAB). Once again, the EAB would consider whether the awarding body concerned had used procedures that were consistent with the
regulatory authorities' code of practice and whether it had done so properly and fairly. Formal application to appeal to the EAB has to be made within three weeks of receipt of the outcome of appeal to the awarding body. Hearings are heard by a panel of three, selected from Board members and from a pool of panellists; they take place within two months of the application and are open to the public.

2.3.2.3.7 Procedures for auditing, monitoring, or evaluating the effectiveness of the review process and for determining ‘customer satisfaction’

Two documents were received in relation to mechanisms for auditing, monitoring or evaluating the effectiveness of EUR services: an evaluation report from OCR which focused upon EUR procedures during 1999; and a report from QCA that presented comparative data, across boards, relating to the processing of EUR requests during 1999. It seems likely that further audits and evaluations were conducted.

2.4 Summary

As noted earlier, the general impression of the intended and actual functioning of the national curriculum review system was positive. Each of the three English EMAs demonstrated a clear commitment to dealing with review requests in a manner that was responsive to the needs of individual schools and consistent with the requirements of the QCA. The general impression from Study 1 was that review procedures were effective, fair and operated consistently between the EMAs.

It should not be concluded, however, that there is no scope for making review procedures more effective or more fair. Indeed, the potential for improvement will be considered at length in the overall evaluation that follows the presentation of results from Studies 1 to 5. Furthermore, there would appear to be scope for increasing consistency.

As the stakes of assessment have increased over the years – and particularly with the advent of assessment for educational accountability – consistency has become the watchword of those charged with responsibility for guaranteeing test and examination standards. The QCA has taken this to heart and the drive for consistency has grounded its most important achievement of recent years: the National Qualifications Framework. As qualifications are formally integrated within this unitary framework, a new impetus arises for resolving differences in procedure that might previously have gone unnoticed or ignored.

The following summary lists the principal procedural differences identified during Study 1. If quantified, the number of points listed might appear to contradict the NFER’s impression of consistency in the processing of review requests. However, many of the
points presented below are relatively minor. Moreover, it was explicitly the intention of the report to focus upon inconsistency rather than consistency; listing the many areas of consistency would have made for dull reading and would have done little to highlight areas where some measure of improvement might be made.

Where differences have been identified it must be considered:

a) whether they are defensible;

b) whether consistency would be preferable;

c) where consistency would be preferable, which procedure should take precedence.

An assessment of the defensibility and desirability of the following differences will be postponed until the overall evaluation.

2.4.1 Inconsistencies between the two key stages

Differences in procedure between the two key stages were few and far between. Indeed, the only major difference appeared to be as follows:

- the R3 review service applied only to key stage 3 English.

2.4.2 Inconsistencies between the three national curriculum test subjects

Differences in procedure, as identified between the national curriculum test subjects, tended to reflect the prevalence of review requests in each subject. In turn, these patterns reflected the different levels of marking reliability that might be expected in subjects with such different marking styles (see Murphy, 1978; 1982; Newton, 1996). The main inconsistencies included:

- different numbers of review markers for English (KS2 ave. = 14; KS3 ave. = 31+), mathematics (KS2 ave. = 2; KS3 ave. = 2) and science (KS2 ave. = 2; KS3 ave. = 4);

- different levels of re-standardisation in different subjects (re-standardisation material was provided only for English);

- different roles adopted by the Lead Chief Markers in different subjects (from the ‘hands-on’ consultative role of the KS3 English LCM to the non-involvement of the KS2 mathematics LCM);

nfer 38
• different provision of standard response templates in different subjects (material was only provided for KS2 science).

### 2.4.3 Inconsistencies between the three English EMAs

Once again, the general impression from Study 1 was that review procedures were effective, fair and operated consistently between the three English EMAs. Yet, a number of inconsistencies were apparent. While there may be much to be said for allowing flexibility in the application of review procedures, it will be necessary to consider which, if any, of the following differences would be defensible, if continued into 2001:

• different management structures (AQA-S subject-based lead officer structure; Edexcel and AQA-N single lead officer structure);

• different tracking forms (forms were not generic across EMAs);

• different instructions to markers (instructions were not generic across EMAs);

• different clerical checking models (AQA-S double check; AQA-N single check plus sample; Edexcel single check);

• different acknowledgement letters (letter formats were not generic across EMAs);

• different interpretations of ‘borderline’ cases (re sufficiency of explanation, provision of deposit cheque, lateness of request, etc.);

• different procedures for requesting missing items (Edexcel fax; AQA-S and AQA-N letters);

• different methods for documenting divergence from intended practice (e.g., Edexcel and AQA-N central list approach; AQA-S individual tracking form approach);

• different levels of provision of desk instructions for staff;

• different policies on the use of a single re-marker for a subject (AQA-S no; Edexcel yes; AQA-N yes with sample check);

• different policies on the sample monitoring of remotely marked work (AQA-N yes; Edexcel no);

• different policies on the use of panel meetings (AQA-N only for English; AQA-S for all subjects; Edexcel only for KS3 English);
• different policies on the sample monitoring of work completed after panel meetings (AQA-N yes; AQA-S no);

• different policies on R3 sample decision making (Edexcel and AQA-N markers recommended decisions to CM and DCM; AQA-S decisions were considered by CM and DCM with review manager);

• different interpretations of scope for group review (and method applied) in subjects other than KS3 English;

• different action following apparent level changes within an R3 sample where the full re-mark was denied (Edexcel automatically amended results; AQA-N and AQA-S did not automatically amend but invited R2 submissions);

• different policies on giving cover letters from schools to markers (AQA-S no; Edexcel and AQA-N yes);

• different levels of manager input into outcome letters and reports;

• different policies on the despatch of R2 and R3 outcome notifications (Edexcel despatched immediately; AQA-N and AQA-S despatched, with scripts, at the beginning of the autumn term);

• different kinds of discussion of the review process in the annual evaluation reports from EMAs.

2.4.4 Differences between the EMAs and the examination boards

The final analysis considered differences between national curriculum review procedures and procedures for processing GCSE/A/AS enquiry-upon-results. Although the two assessment systems are significantly different, there are good reasons for making such comparisons. Firstly, examination boards have been successfully processing EUR requests for many years, so there may be important lessons to be learned from procedural differences. Secondly, both assessment systems are under the control of the QCA, either directly or indirectly through regulation. In fact, Section 7 of the QCA's code of practice (QCA, 2000a) is quite explicit in discussing how EURs should be processed. As such, it would be important to highlight any principles manifest in one system that were contradicted by principles manifest in the other. In this spirit, the following differences are worth considering:

• GCSE/A/AS external marking procedures are externally regulated (while national curriculum external marking procedures are not);
the principles underlying GCSE/A/AS external marking procedures are laid down in a code of practice which is in the public domain (while principles underlying national curriculum external marking procedures are less explicitly recorded in contracts with EMAs that are not in the public domain);

in addition to guidance for schools on EUR services, the Joint Council for General Qualifications publishes guidance for re-markers and awarding body staff (while the QCA, through the EMAs, publishes only guidance for schools);

at GCSE/A/AS, the emphasis is upon explaining to schools why grade change occurred during EUR (in contrast to the national curriculum context, in which the emphasis is upon explaining to schools why level change did not occur during review);

EUR procedures tend to be couched in more authoritative language than review procedures;

EUR procedures generally involve either full clerical re-checks or full re-marks (while review procedures generally involve partial re-marking);

different awarding bodies are allowed to offer different EUR services (while EMAs are required to be consistent);

EUR services do not necessarily include a group review option (while review services do, but only for key stage 3 English);

when a group review EUR service is offered, the initial sample is drawn by the awarding body (while R3 review samples are drawn by schools);

awarding bodies did not generally down-grade candidates during 2000 (whereas, EMAs explicitly warn that levels can go up or down);

awarding body procedures require the extension of re-marking (even to schools that have not submitted EURs) when suspect initial marking is identified (while EMA procedures tend not to require this);

information booklets from awarding bodies concern only post-examination services and are distributed with results (while EMA booklets concern marking generally and arrive much earlier);

the awarding bodies make use of their web-sites to explain EUR services (while the EMAs do not);
• the awarding bodies charge higher fees for EUR services (than EMAs do for review services);

• different EUR fees are charged by different awarding bodies (while EMA deposits are constant);

• awarding bodies charge for clerical re-checks (while EMAs do not);

• EUR fees are invoiced (while review deposits are required in advance);

• requesting an EUR service requires less form-filling (than requesting a review);

• centres have more time to register an EUR request (than schools have to register a review request);

• EUR re-marks are carried out in black ink (while review re-markers use green);

• EURs that result in mark changes are recorded on MJCR forms which are for internal use only (while reviews that do not result in mark changes are explained on forms for external dissemination);

• EUR services require that requests be processed on a first-come, first-served basis (while review services tend to prioritise R1 reviews);

• there was no evidence of monitoring of EUR re-marking (while there was some evidence of monitoring of review re-marking);

• there are published turn-around times for EUR services (but not for review services);

• EUR services specify that fees will be returned only for candidates whose grades have changed (whereas review deposits are returned even if only one pupil from a school’s submission is awarded a higher level);

• there are formal internal and external appeal mechanisms for EUR services (while there are none for review services).
Section 3 An analysis of data relating to the external marking review services (Study 2)

3.1 Introduction

The intention behind Study 2 was to collate and represent quantitative evidence that might support the evaluation of the external marking review services during 2000. In addition to considering a range of national summary data for 2000, this would involve exploring patterns in review requests and outcomes between review types, key stages, subjects, External Marking Agencies and over time.

Inevitably, even if clear patterns or differences were to be detected, their interpretation would not be straightforward or unambiguous. Structural disparities ensure the complexity of this interpretative task. For example, differences in the number of review requests between key stages might reflect a range of differences between primary and secondary contexts, such as the impact of performance tables at key stage 2 or the larger cohort sizes at key stage 3. Similarly, interpreting changes in the number of review requests over time would be complicated by the way in which EMA responsibilities have changed and might also be affected by changes in the way in which review request data had been collected and represented over time.

Nevertheless, the collation and representation of quantitative data is an essential component of the present evaluation project and, as long as essential caveats are borne in mind, constitutes a vital source of evidence.

3.2 Methodology

To explore a full range of issues, it was deemed important to collate a range of information from a variety of sources. These were to include the following:

1. aggregate school- and pupil-level data, provided by the QCA, relating to the number of review requests over time and the outcomes of these review requests;

2. individual school- and pupil-level data, provided by the EMAs, relating to the number of review requests during 2000 and the outcomes of these review requests;

3. aggregate school- and pupil-level data, provided by the QCA, relating to GCSE/A/AS-enquiry-upon-results (EUR) services during 1999 (to provide an alternative perspective from which to consider the national curriculum review data).
3.2.1 Aggregate review request data

The aggregate review request data were provided by the QCA in the form of overall summary reports and supplementary tables; the tables contained finer grained analyses, from which the overall summary reports were produced. The overall summary reports contained aggregate review request and review outcome figures that had been (or that were to be) released to the public. These presented results by review type and by subject but not by EMA. Data relating to the EMAs were provided by the QCA in the additional tables that were not available to the public.

Each year, aggregate data are collected directly from the EMAs through a Management Report reporting mechanism. At regular intervals (typically weekly), throughout the duration of the review process, each EMA sends to the QCA an aggregate school- and pupil-level Management Report for each key stage. These list the number of review requests received (by subject and by review type), the number of requests which have so far been processed, the number of these processed requests for which levels have been changed, and the number of requests that remain unprocessed. Within each EMA a central database is maintained from which these Management Reports are generated.

At the end of the review process, when no review cases remain outstanding, the QCA collates the final Management Reports from the three EMAs to provide summary data at the national level.

3.2.2 Individual review request data

As the data provided by the QCA were at an aggregate level, they did not present the level of detail that was necessary for a more thorough analysis. To conduct a more detailed exploration of data for 2000 it was essential to interrogate directly the review databases maintained by the three EMAs. In the Schools’ Guide for 2000 it was explained that all review requests would be completed by Friday 22nd September (and that schools had until Saturday 30th September to report these results to parents or guardians). A request was made to each EMA to provide the NFER with copies of the following databases by Friday 29th September:

1. a pupil-level database with fields including key stage, subject, review type, initial level and final level;

2. a school-level database with fields including key stage, subject, review type and outcome return date.
As AQA maintained a central database, the records for AQA-N and AQA-S, respectively, were provided by AQA-S.

As data from the EMA databases were used directly to produce the QCA overall summary reports there should, in theory, have been a direct mapping between results that were presented in the aggregate data for 2000 and those that emerged from analysis of the pupil- and school-level EMA databases.

In fact, while the results were clearly 'in the same ball-park', there were slight differences. This was primarily due to two factors: firstly, the databases that were received by the NFER did not have final outcome levels for all pupils (although these were present for the vast majority of pupils from all EMAs). While the QCA resolved such issues through manual reconciliation, these data were omitted from the analyses conducted by NFER. Secondly, while preparing the EMA databases for analysis, the NFER recoded a number of records. For instance, in the small number of cases where either the initial level or the final level column had a two-figure digit in it, this was assumed to be an input error and was recoded as the first digit only. More significantly, where either the initial level or final level was a letter rather than a number (i.e., A, B, D, L, V, M, X) these were excluded from the analyses and explored separately. Full details of these adjustments can be found in Appendix 3.1.

Finally, the pupil-level EMA databases containing outcome data were somewhat ambiguous for key stage 2 English. Because key stage 2 performance tables publish school-level results for the Reading and Writing tests in addition to the overall result for English, it had been agreed that schools should be able to submit review requests where a pupil's level in Reading, Writing, English or any combination of these might change. Unfortunately, this creates a complicated classification problem: if a request is made for a review in Writing (only), but the re-mark also leads to a change in level for English overall, does that mean that the pupil should be counted as having made a successful request in English? Similarly, if a request is made for a review in English (only), but a clerical check of the Writing paper leads to a level change for Writing but not for English, should that be counted as a successful request in Writing or a failed request in English? These issues appear not to have been resolved. At the time of writing the present report it was not clear exactly what classification systems had been applied by AQA-N or AQA-S (or even whether they had applied the same system). The database that was received from Edexcel presented initial and final levels for Reading, Writing and English separately, thereby leaving the classification to NFER. The system applied is explained in Appendix 3.1.
3.2.3 Aggregate EUR request data

To provide some contextualisation for the review data, data from EUR services from 1999 were also collected and represented. As for the review services data, aggregate EUR data were provided by the QCA in the form of overall summary reports and supplementary tables (which contained finer grained analyses, from which the overall summary reports were produced). The overall summary reports contained the aggregate EUR data that were released to the public (QCA, 2000b). These presented information by EUR type and by awarding body but did not present data by subject. Data relating to the subjects were provided in additional tables which were not in the public domain. Data relating to the summer 2000 examination series would not have been ready for publication until February 2001, which is why only 1999 data were collected.

3.3 Results

The results will be presented in three sections, corresponding to the three different data sources. This begins with an exploration of the summary data provided by QCA.

3.3.1 Aggregate review request data

From aggregate data provided by the QCA, a range of charts and table were produced. These illustrate changes over time in the prevalence and nature of review requests and review outcomes, broken down by review type, key stage, subject and BMA. The charts are presented within the main body of the report while the more detailed tables are to be found in Appendix 3.2.

3.3.1.1 The number of schools requesting reviews

Data concerning the number of schools that lodged review requests are represented in Figures 1 to 7 and in Tables 1 to 22. Figures 1 to 3 illustrate the number of requests lodged at KS2 by subject and by review type. In all, 838 schools requested R1 reviews for KS2 English during 2000, 476 for maths and 344 for science (Table 1). This represented 6%, 3% and 2% of all schools in England that participated in KS2 testing during 2000 (Table 8). Correspondingly, 1343 schools requested R2 reviews for KS2 English, 487 for maths and 536 for science (Table 2). This represented 9%, 3% and 4% of all participating schools (Table 9). Thus, around twice as many schools requested R1 and R2 reviews for English as were requested for maths and science. During 1998 and 1999 a similar pattern held, although the number of schools that requested R2 reviews for KS2 science during 1998 was inflated. Generally speaking, the number of schools requesting reviews over time seems to have remained fairly stable between 1999 and 2000.
Figure 1. Number of schools that requested R1 reviews at KS2

Figure 2. Number of schools that requested R2 reviews at KS2

Figure 3. Number of schools that requested any reviews at KS2
Figure 4. Number of schools that requested R1 reviews at KS3

Figure 5. Number of schools that requested R2 reviews at KS3

Figure 6. Number of schools that requested R3 review at KS3

Figure 7. Number of schools that requested any reviews at KS3
When the numbers of review requests from schools were explored by EMA, it appeared that trends between subjects as well as absolute values, were very similar (see, for example, Tables 8 to 10 which present figures representing the percentage of the total number of participating schools in each EMA's catchment, meaning that they are directly comparable between EMAs).

Figures 4 to 6 illustrate the number of requests lodged at KS3 by subject and by review type. In total, 346 schools requested R1 reviews for KS3 English during 2000, 408 for maths and 523 for science (Table 11). This represented 10%, 10% and 13% of all schools in England that participated in KS3 testing during 2000 (Table 19). While this is higher than the percentage of primary schools that requested R1 reviews it must be recalled that there are only a third the number of secondary schools as primary schools (and, therefore, correspondingly more pupils per cohort). Regarding the R2 review requests, these were received from 482 schools for English, 162 for maths and 350 for science (Table 12). This represented 14%, 4% and 9% of the cohort, respectively (Table 20). Finally, 402 schools (11% of those participating) requested R3 reviews for KS3 English (Table 13/21).

With the R3 schools factored in (Figure 7), it was clear that a similar pattern was observed between subjects at KS3 as was observed at KS2, with the number of schools requesting English reviews far outstripping that for maths and science. At KS3, however, there appeared to be more schools requesting reviews for science than for maths. This was a pattern that had been replicated since 1998 (at least).

As at KS2, when these patterns were explored by EMA, it appeared that trends between subjects as well as absolute values, were very similar (Tables 15 to 22). The most notable divergence was for science at Edexcel, where the percentage of schools that requested any review was 10% higher than for AQA-S or AQA-N (Table 22).

Considering patterns over time at KS3, there was an overall increasing trend in the number of schools that had requested any review since 1998 that was not apparent at KS2 (Figure 7 versus Figure 3). Indeed, reflecting increases for R1 and particularly for R2 (but not as clearly for R3) there appeared to have been a steady increase in the number of secondary schools requesting reviews in English since 1998. For maths and science, however, the overall figures for 2000 were similar to those for 1999.
Figure 8. No. KS2 pupils for whom R1 reviews were requested

Figure 9. No. KS2 pupils for whom R2 reviews were requested
Figure 10. No. KS3 pupils for whom R1 reviews were requested

Figure 11. No. KS3 pupils for whom R2 reviews were requested
3.3.1.2 The number of pupils for whom reviews were requested

Turning now to the number of pupils that were involved in schools' review requests, the trends since 1997 are presented in Figures 8 to 11. Data have been represented for R1 and R2 reviews, but not for R3. This is because there was inconsistent practice between EMAs in the way in which this data was recorded (if at all).

A quick inspection of Figures 8 to 11 reveals that there are few clear consistent trends over time in the number of pupils involved in review requests in the three subjects at each key stage.

The closest to an observable trend was apparent for R2 reviews in English: at both key stages, the number of review requests has increased steadily over time. For maths, at both key stages, the tendency has been for a consistent number of R2 review requests from one year to the next. Science has witnessed a similar trend, although punctuated by odd years during which considerable increases have been observed.

Turning to R1 reviews, far less in the way of consistency over time was observed. Changes appear to have been relatively erratic for all subjects at both key stages. Perhaps the most notable change was for English at KS2, which rose to a particularly high level during 1999 and stayed high during 2000.

If the number of pupils involved in review requests is divided by the number of schools submitting review requests, this gives an indication of the average number of pupils submitted by each school. At KS2, these figures fluctuated at around 1½ to 2 for all subjects and both review types, with the exception of English R2 which, in 2000, had an average of 4½ (Table 25 and 26). At KS3, the figures fluctuated at around 2 to 2½ for all subjects except KS3 science and KS3 English which, in 2000, had respective averages of 3½ and 8 (Table 30 and 31).

Overall, the percentage of the cohort involved in review submissions ranged from 0.05% for KS3 maths R2 to 0.99% for KS2 English R2.

3.3.1.3 The number of pupils whose levels were changed following review

Overall, the percentage of the cohort whose levels changed following review ranged from 0.03% for KS3 maths R2, through 0.46% for KS3 English R2 (0.47% for KS2 English R2), to 2.50% for KS3 English R3. Data for R1 and R2 reviews are presented in Figures 12 to 15 and in Tables 33 to 42 (which also include data for R3).
Figure 12. Percentage of KS2 pupils, for whom R1 review requests were made, whose levels were changed

Figure 13. Percentage of KS2 pupils, for whom R2 review requests were made, whose levels were changed
Figure 14. Percentage of KS3 pupils, for whom R1 review requests were made, whose levels were changed

Figure 15. Percentage of KS3 pupils, for whom R2 review requests were made, whose levels were changed
Generally speaking, the figures illustrate a reasonable level of stability between 1999 and 2000. For R1 reviews during 2000, at both key stages, the probability of a successful review outcome was high for most subjects, ranging between 87% for KS2 science and KS3 English to 95% for KS3 maths and science (Table 35 and 40). The one slightly odd exception was for KS2 English R1 reviews, where the percentage that resulted in a level change fell from 75% in 1999 to 58% in 2000. This discrepancy will be returned to in 3.3.2.2.

The figures for R2 reviews also exhibited a reasonable level of stability between 1999 and 2000, although there was a general trend towards a slightly higher percentage of pupils with changed levels during 2000. This was most marked at KS3 and for KS3 science in particular.

Interestingly, for R2 reviews (R1 to a lesser extent) there was a consistently higher percentage of pupils with changed levels at KS3 when compared with the same subject at KS2 (for both 1999 and 2000). Indeed, during 2000, the highest percentage of pupils with changed levels at KS2 (55% for maths) was lower than the lowest percentage of pupils with changed levels at KS3 (62% for science).

### 3.3.2 Individual review request data

Requesting pupil- and school-level databases from the EMAs meant that subtler analyses could be completed. Three principal analyses were conducted:

1. a deeper investigation into the number of review requests lodged by each school;
2. a deeper investigation into the outcomes of review requests for individual pupils;
3. an investigation into the dates by which review outcomes were returned to schools.

#### 3.3.2.1 The number of review requests lodged by each school

The first use to which the pupil-level databases were put was to explore in greater depth the number of review requests submitted by schools in each subject and at each key stage. (This was possible because each pupil record had a DfEE number attached, meaning that pupil-level data could be aggregated to school-level.)

The results of the analyses are presented in Table 43 (for all 3 EMAs combined) and Tables 43a to 43c (for each EMA respectively). It should be noted that these analyses have only been conducted for R2 reviews, but that all pupils submitted for R2 reviews have been included.
The various elements of Table 43 present the frequency and cumulative frequency distributions of the number and percentage of schools that submitted each of a range of numbers of pupils. It was apparent that, across all subjects and both key stages, the modal number of submissions per school was one pupil. That is, the largest number of schools were in the 'submit one pupil' category. However, this does conceal a considerable amount of variation, which the mean number of pupils helps to unravel. Here, as discussed earlier, the average number of pupils submitted ranged from 1½ (KS2 maths) to 8 (KS3 English). Clearly, then, there were a considerable number of schools that submitted a considerable number of pupils. Indeed, as can be inferred from Table 43, over one-quarter of KS3 schools submitted more than 10 pupils for R2 review in English and over one-fifth of KS2 schools submitted more than 5. Very similar patterns were observed across EMAs.

3.3.2.2 The outcomes of review requests for individual pupils

Individual pupils' levels were analysed to explore in more depth the nature of review outcomes. The basic comparison was between pupils' initial levels (before review) and their final levels (after review). To explore the nature and extent of change, pupils with unconventional levels (i.e., letters such as A, D, M, etc.) were excluded from the analyses (see Appendix 3.1). As indicated in Table 47, there were only a small number of cases that fell into this category so the remaining data are still quite representative.

The first analysis, as presented in Table 44 (and 44a to 44c), was simply to categorise each pupil in terms of their change in level following review. In principle, there are only three states: to remain with the same level; to be awarded a higher level; or to be awarded a lower level. In fact, Table 44 broke these down further according to how many levels higher or lower a pupil was awarded.

The first point to be aware of is that levels that had been input by the EMA as zero or one were left as such, regardless of whether this was actually a valid level to have been awarded. Similarly, candidates that had been awarded a level N were recoded to 1 prior to analysis. These conventions may well explain many of the large changes evident from Table 44.

The most important and revealing data are found towards the bottom of Table 44, where pupils were categorised as simply 'down', 'up' or 'same'. In fact, the table in the bottom-left corner of Table 44 (data aggregated across EMAs) is an analogue of the 2000 data from Tables 35, 36, 40 and 41. Reassuringly, the percentages of pupils whose levels changed following review were almost identical between the QCA aggregate and the EMA pupil-level sources. The one exception was the odd-looking figure of 58% for KS2.
English R1 which, on the pupil-level analysis, was 73%. Supported by the fact that the corresponding QCA aggregate figure for 1999 was 75%, it may be that the QCA aggregate figure for 2000 was not correct (and, therefore, that there had been no fall in the percentage of pupils whose levels were changed following R1 review at KS2).

Table 44s (below) is a summary table presenting comparative data across EMAs. It indicates that, for six of the subject-by-type-by-stage comparisons, significant differences ($P < 0.05$) were observed between EMAs in the percentages of review requests that resulted in level changes.

**Table 44s. The percentage of review requests that resulted in level changes.**

<table>
<thead>
<tr>
<th></th>
<th>AQA-N</th>
<th>AQA-S</th>
<th>Edexcel</th>
<th>Chi$^2$ probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 E R1</td>
<td>57</td>
<td>64</td>
<td>89</td>
<td>$&lt;0.01$</td>
</tr>
<tr>
<td>2 M R1</td>
<td>93</td>
<td>91</td>
<td>95</td>
<td>0.26</td>
</tr>
<tr>
<td>2 S R1</td>
<td>89</td>
<td>81</td>
<td>94</td>
<td>$&lt;0.01$</td>
</tr>
<tr>
<td>2 E R2</td>
<td>36</td>
<td>50</td>
<td>43</td>
<td>$&lt;0.01$</td>
</tr>
<tr>
<td>2 M R2</td>
<td>51</td>
<td>52</td>
<td>61</td>
<td>0.06</td>
</tr>
<tr>
<td>2 S R2</td>
<td>48</td>
<td>53</td>
<td>49</td>
<td>0.43</td>
</tr>
<tr>
<td>3 E R1</td>
<td>86</td>
<td>81</td>
<td>95</td>
<td>$&lt;0.01$</td>
</tr>
<tr>
<td>3 M R1</td>
<td>92</td>
<td>96</td>
<td>96</td>
<td>0.04</td>
</tr>
<tr>
<td>3 S R1</td>
<td>93</td>
<td>96</td>
<td>96</td>
<td>0.20</td>
</tr>
<tr>
<td>3 E R2</td>
<td>74</td>
<td>72</td>
<td>62</td>
<td>$&lt;0.01$</td>
</tr>
<tr>
<td>3 M R2</td>
<td>77</td>
<td>75</td>
<td>76</td>
<td>0.92</td>
</tr>
<tr>
<td>3 S R2</td>
<td>66</td>
<td>61</td>
<td>61</td>
<td>0.28</td>
</tr>
</tbody>
</table>

In fact, Table 44s adds a further twist to the issue of the anomalous R1 review figure for English at KS2. The first row of Table 44s represents the largest of the observed differences between EMAs, which stretches from just below the QCA aggregate figure (58%) to well above the EMA pupil-level figure (73%).

Another important issue arising from Table 44 was the percentage of pupils whose levels were lowered following review. Across key stages, between 13% (KS2 English) and 18% (KS3 science) of all R1 review requests resulted in the award of a lower level. This contrasted sharply with the percentage of R1 review requests that resulted in the award of a higher level, which ranged between 60% (KS2 English) and 79% (KS3 maths).
Figure 16. All R2 review outcomes at KS2

Figure 17. All R2 review outcomes at KS3
An even more extreme pattern was observed for R2 review requests. The largest percentage of R2 review requests lowered was 1% (for both KS2 English and KS3 maths) while the most frequent percentage lowered was 0%. Yet, the percentages of pupils whose levels were raised ranged from 43% (KS2 English) to 75% (KS3 maths).

Table 45 and 46 present outcome data by the initial level awarded to each pupil. Table 45 includes outcomes for R1, R2 and R3 together. Table 46 includes only outcomes for R2 review, but also includes an analysis by subject. Figures 16 and 17 represent the KS2 and KS3 outcome data relating to R2 reviews for all subjects combined (see Table 46).

The figures highlight a clear trend at KS2 which was somewhat less pronounced at KS3. At KS2, the probability of being awarded a higher level appeared to decrease as the level initially awarded to the script increased. Thus, while 55% of level 3 pupils’ scripts were awarded a higher level following re-mark, this occurred for only 43% of level 4 pupils and only 3% of level 5 pupils (an increase from level 5 to level 6 would probably imply a review of extension test scripts). This pattern was similar across all EMAs. At KS3 the trend was less dramatic, with 72% of level 4 pupils being awarded a higher level, 70% of level 5 pupils, 59% of level 6 pupils and 54% of level 7 pupils (an increase from level 7 to level 8 would probably imply a review of extension test scripts).

Particularly at KS2, the combined figures were swamped by English review requests. Yet similar, although somewhat less extreme, trends were apparent for maths and science at KS2, where three-fifths of level 5 pupils’ scripts remained at the same level following review compared with fewer than half of level 4 pupils’ scripts.

3.3.2.3 The dates by which review outcomes were returned to schools

The school-level databases were used primarily to establish the extent to which review requests were processed on time. As explained earlier, EMAs should have processed all review requests by Friday 22nd September 2000. Tables 48, 49 and 50 present the number, percentage and cumulative percentage (respectively) of review requests returned by the end of each week during the review period. In principle, the date field used to produce Tables 48 to 50 was supposed to represent, for each school, the date when their review outcome was despatched by their EMA. This was not necessarily the date by which a school’s review submission had been completed. In fact, EMAs typically retained reviews completed during the summer break and despatched outcomes at the beginning of the autumn term. Unfortunately, it appears that there were inconsistencies in the ways in which these data were entered, both between and within EMAs. For example, prior to the 8th September, the Edexcel data apparently represent review completion date rather than review despatch date (there was a mass despatch of completed reviews on the
Figure 18. Cumulative percentage of R1 outcomes returned by each week

Figure 19. Cumulative percentage of R2 outcomes returned by each week

Figure 20. Cumulative percentage of R3 outcomes returned by each week
8th). Yet, after the 8th, the data apparently represent despatch date. Such inconsistencies will inevitably confuse the interpretation of the evidence.

Overall, 99% of English, maths and science R1 review outcomes had been returned to schools by the 22nd September. Of the R2 review outcomes, 99% of maths and 98% of English and science outcomes had been returned by the 22nd. Finally, 57% of the R3 review outcomes had been returned to schools by the 22nd (see Table 50). It was not until October the 6th that more than 95% of R3 review outcomes had been despatched.

Figures 18 to 20 present the return-date data, aggregated across subjects, by review type and EMA. Figure 18 shows that, for all EMAs, the vast majority of R1 reviews were completed and despatched by Friday 4th August. From Figure 19, AQA-N appears to have completed and returned the vast majority of R2 reviews by 18th August. AQA-S and Edexcel appear to have completed and despatched a steady number throughout the summer break until the week ending 8th September, by which time all EMAs had completed and returned the vast majority of R2 reviews (although this picture will be misleading if at least some of the data actually represented completion date rather than despatch date). Finally, Figure 3 shows that the EMAs tended to despatch the majority of completed R3 reviews within the space of one or two weeks; AQA-N first, then Edexcel, then AQA-S. AQA-S and Edexcel appear to have taken longer than AQA-N to despatch the tail-end of the R3 reviews.

3.3.3 Aggregate EUR request data

Tables 51 to 56 represent the aggregate EUR data, for GCSE and GCE, from 1999; first by subject, then by service type, then by awarding body. As the EUR system is so different from the review system (see Study 1), both in principle and in practice, it would not be appropriate to draw any direct comparisons between the resultant data. The EUR data are presented simply to provide another perspective from which to consider the review data.

Table 51 and Table 52 present EUR summary data for a selection of English, maths and science subjects at GCSE and A/AS. At GCSE the number of enquiries as a percentage of the total number of candidate entries ranged from 0.11% (statistics) to 0.82% (biology); at A/AS, the percentages ranged from 1.10% (other science) to 6.77% (English). Only at A/AS was there a tendency for English to receive more EUR requests than the other subjects.
Figure 21. Percentage of GCSE candidates, for whom enquiries were lodged, whose grades were changed

Figure 22. Percentage of GCE candidates, for whom enquiries were lodged, whose grades were changed
Figures 21 and 22 represent the percentage of GCSE and GCE candidates, for whom EUR requests were made, whose grades were changed. For most of the subjects at A/AS, the percentage of candidates whose grades were raised ranged between 8% and 17%. At GCSE, the percentages with changed levels tended to range between 3% and 13% for the sciences and to fall at around 21% for English (although the data for statistics appear to be higher than for any other subject, only a few candidates were submitted for EUR).

When the EUR results were broken down by service type (Table 53 and 54) there were few patterns of particular significance. At A/AS, only a few clerical re-checks resulted in changed grades (8%), fewer than for the re-marking services (17%). At GCSE, similar percentages of candidates received changed grades following clerical re-check and re-marking (around 19%).

Finally, Table 55 and 56 present turn-around times by awarding body. The awarding bodies are supposed to meet a 40-day turn-around target for enquiry services (30-day for priority services). During 1999, CCEA and WJEC fared best in meeting these targets. Edexcel tended to fare worst, with only a tiny minority of enquiries processed within specified targets for most services other than the clerical re-checks.

3.4 Discussion

Before turning to an evaluation of results that emerged from Study 2, a few words on the quality of the data sources are in order. The integrity of data arising from the national curriculum review system relies principally upon the integrity of the EMA databases and the care with which these are maintained. There is clearly work to be done, between the QCA and the EMAs, to ensure that procedures for recording and reproducing essential review data are tightened.

3.4.1 Quality of review data

Although a number of important issues of quality will be addressed in this section, most attention will be devoted to the particular problem of key stage 2 English, which very much complicated the collation of national data during 2000.

3.4.1.1 Classifying key stage 2 English request and outcome data

Perhaps the most awkward data issue that needs to be resolved for 2001 is the classification of review requests for KS2 English. As discussed earlier, this is particularly complicated because it is possible for one pupil to be classified simultaneously into more than one category. For example, a school may submit a review request for Writing that it
believes will not result in a change of level for English overall. If this did turn out to be
the outcome of the review, then it could be classified in a number of ways, for instance:

1. as one review request (E) and no level change;

2. as one review request (W) and one level change (W);

3. as two separate review requests (W & E) and one level change (W).

The first two of these classification models are motivated by the desire not to double
count review requests; they require that all key stage 2 English review requests, relating to
the same pupil, are recorded either as requests for Writing or as requests for English.

Adopting the first model, all requests for Writing (or Reading) would be counted simply
as requests for English. The problem with this is that there may have been no expectation
(on behalf of the submitting school) that the Writing request would have led to a change
of level for English. Indeed, if the Writing review had led to a higher grade for Writing
but not for English, then it would have been classified (under the first model) as a failed
English request rather than as a successful Writing request. Two key obstacles are
associated with the first approach; firstly, that the overall outcome results for English will
not be comparable with those for maths or science (as many schools will have submitted
requests for Writing or Reading with no expectation of an overall level change for
English); secondly, there will be no separate review request or outcome data for Writing
or Reading individually.

Adopting the second model, all requests for Writing (or Reading) would be counted simply
as requests for Writing (or Reading). While this means that there would be
separate review request and outcome data for Writing (and Reading), it also means that
any Writing requests that resulted in changes for English overall, would not be counted as
overall English changes. Once again, then, the overall outcome results for English would
not be comparable with those for maths or science. Moreover, there would be a further
problem of how to classify requests that involved both Writing and Reading (without
double counting).

Unfortunately, adopting the third model would not resolve a great deal more. As for the
second model, it would ensure that all Writing (or Reading) requests were recognised as
such in final reports. Moreover, there would be no problem if a pupil’s Reading and
Writing scripts were both subject to review as these could be counted separately.
However, the data for English overall would still not be comparable with those for maths
or science (as, once again, they would include those schools that submitted requests for
Writing or Reading with no expectation of an overall English level change).
It should be noted that these three models are actually constructed from first principles, ignoring the actual process by which review requests are lodged. In fact, Form R2 for key stage 2 English (the submission form at the back of the Schools' Guide) clearly puts the classificatory onus upon schools. The school has to complete ‘current level’ and ‘requested level’ boxes for English, Reading and Writing separately (for each individual pupil). Where ‘requested’ is higher than ‘current’ it must be assumed that the school has requested a review. This supports the third model where pupils are explicitly double (or triple) counted, where necessary. Thus, English would include some but not all Reading or Writing requests, i.e., only those that the school believed would also lead to an overall change of level for the subject. Moreover, and for this reason, this model would help to ensure that results for English were comparable with those for maths and science. Unfortunately, however, this model is not fool-proof. For example, it will still fail to recognise reviews that are submitted by the school as only Writing (or Reading) but that turn out also to result in changes for English overall.

Finally, there is an alternative interpretation of the double counting model that leaves the classification of Writing (and Reading) unchanged but that revisits the classification for English. Under this final model Writing (and Reading) would be classified precisely as described above. However, and in addition, any review request at all (Writing, Reading, English or any combination) would be classified as a request for English and any change in level at all would be classified as an outcome change for English (whether for Writing, Reading, English or any combination).

Although, no doubt, there are other alternative models, there is no need to dwell further on the issue here. There are two points to make; firstly, QCA must be precise and explicit concerning what model it expects the EMAs to be using; secondly, when data are published, it must be made explicit exactly how they should be interpreted, particularly stressing any differences between subjects.

3.4.1.2 Additional data quality issues

Three further issues arose in exploring the review data for 2000, the first concerning R3 reviews, the second relating to data specification issues and the third regarding general data management.

First, there is a serious ambiguity in the interpretation of the number of pupils involved in R3 review requests. From a principled standpoint it is not clear whether this figure ought to represent the total number of pupils, from each school, who took the English test; or whether it should represent the total number of pupils in the group which was submitted to the EMA as part of each school’s R3 request. Assuming the latter, practical problems
clearly arose during 2000. Although schools are supposed to record the total number of pupils that they submit on their R3 request forms, the EMA databases calculated the number of R3 pupils from the number for whom review outcome data was actually entered onto their systems. Where R3 reviews were rejected, pupils’ details were typically not recorded; indeed, AQA appeared only to have entered pupils’ details if pupils’ levels changed following R3 review. A decision needs to be taken concerning how to address this issue for 2001.

Second, it was clear from speaking to QCA managers that they had expectations of the kind of review data that were being routinely collected by the EMAs that were not, in fact, being consistently recorded. In particular, review turn-around times could not be evaluated effectively as different EMAs recorded different start and finish dates, corresponding to different critical stages. The clearer specification of to-be-recorded data, especially salient dates, is an issue that should be reconsidered for 2001.

Finally, the databases received by NFER from the EMAs contained a number of input errors. While the number of these was very small, it begs the question of whether data management systems could be improved (for example, with the addition of input-masks).

3.4.2 Overall impressions from the review data

We can now turn to the substantive findings that emerged from Study 2; we begin with the national summary data for 2000 before turning to a more detailed investigation into underlying patterns and trends.

Nationally, no more than 1 in 100 pupils were submitted for key stage 2 reviews in any of the subjects for either R1 or R2 during 2000. At key stage 3, the comparable figure was fewer than 1 in 100 pupils. Interestingly, the number of GCSE enquiries as a percentage of GCSE entries also emerged as fewer than 1 in 100 pupils (1999 data), while the comparable figures for A-level tended to range between 2 and 7 pupils per 100. Unfortunately, there are so many structural differences between the systems, let alone differences in the ways in which these data were compiled, that it would be folly to read too much into such direct comparisons. Perhaps the most that might be inferred is that the national curriculum review request figures were not obviously ‘in the wrong ball-park’ in comparison with those from its nearest relative.

A closer interrogation of the review data revealed that, of all schools participating in national curriculum testing at key stage 2 during 2000, 6% (English), 3% (maths) and 2% (science) requested R1 reviews for at least one pupil. In fact, the mean numbers of pupils for whom schools requested R1 reviews were 2 (English), 1½ (maths) and 1¼ (science).
The comparable percentages for R2 reviews were 9% (English), 3% (maths) and 4% (science), while the comparable means were 4½ (English), 1½ (maths) and 2 (science).

At key stage 3 during 2000, of all schools that participated in national curriculum testing, 10% (English), 10% (maths) and 13% (science) requested R1 reviews for at least one pupil. The mean number of pupils for whom schools requested R1 reviews was 2 (English), 2½ (maths) and 2 (science). The comparable percentages for R2 reviews were 14% (English), 4% (maths) and 9% (science), while the comparable means were 8 (English), 1½ (maths) and 3½ (science).

What is clear from these data is that, while the numbers of pupils involved in review requests were relatively small, the numbers of schools involved were sizeable. While this was entirely as might have been expected — if marking errors had been distributed more randomly than systematically — it is not without trivial implication. Particularly at key stage 3, where one-third of schools submitted some kind of review request for English and one-fifth for science, the profile of the review system will inevitably be high. Moreover, with large percentages of all review requests resulting in level changes, the review system will present large numbers of schools with formal confirmation of their concerns regarding the quality of the initial marking of scripts.

When review outcomes were explored in more depth it was interesting to note the number of review requests that led to lower levels (data that is released into the public domain is classified simply as ‘same’ or ‘change’ and no distinction is made between outcomes that go up or down). The foremost conclusion was that far fewer review requests led to levels being lowered than to levels being raised. For R1 reviews the ratio was of the order 1:5. Indeed, virtually no R2 reviews led to pupils’ levels being lowered. If one were to assume that marking errors resulted in pupils being awarded too high a level as frequently as they were awarded too low a level, then this finding suggests that there were substantial numbers of marking errors that were not rectified during the review period (and that, presumably, were simply accepted by schools). Although not investigated directly, this must predominantly be a consequence of far fewer reviews being submitted to rectify errors that resulted in levels that were too high. To the extent that the review system is ineffective in rectifying certain kinds of marking error (i.e., those that lead to inappropriately high levels) this emphasises the importance of removing as much error as possible during the initial marking phase.

Another interesting finding was the interaction between initial level awarded to a script and the probability of a re-mark resulting in a level change. This was particularly salient at key stage 2, where only 3% of the 648 re-marked scripts at level 5 were awarded a higher level. It is presumed these data related predominantly (if not exclusively) to the re-
marking of extension test scripts for pupils who had received insufficient marks to be awarded level 6. Although only a small percentage of key stage 2 pupils are ever awarded level 6 from performance on the extension test, the very high rate of failed review requests is still worth highlighting. It might suggest, for example, that teachers did not find the extension test mark schemes sufficiently accessible; alternatively, it might be a consequence of teachers not consulting the mark schemes at all before submitting review requests.

Even on average, though, only 45% of re-marked scripts at key stage 2 resulted in a level change. This seems to be an unexpectedly low rate, given that schools had access to both scripts and mark schemes. Even at GCSE and A/AS, where scripts were not available to schools during 1999, success rates following EUR were of the order of around 10% (i.e., Biology, Chemistry and Physics at GCSE and GCE) to 20% (i.e., English at GCSE).

Finally, when return dates were analysed, it became clear that the vast majority of R1 and R2 review outcomes were despatched to schools by the end of the 22nd September 2000 – the date by which EMAs were supposed to have completed all review requests. This would appear to be satisfactory, on the whole. However, only 57% of R3 review outcomes were returned by the 22nd – a rate that is far less satisfactory (although the differences between EMAs should be noted here).

3.4.3 Patterns that emerged from the review data

To some extent, it was hard to judge the acceptability of data arising from the national curriculum review system, owing to the lack of an external system which could be used as a benchmark. As such, further insights that were to be gained from internal comparisons became even more prominent. This is why an important aim of Study 2 was to explore patterns in review requests and outcomes between: review types (R1, R2 and R3); national curriculum subjects (English, maths and science); key stages (2 and 3); EMAs (AQA-N, Edexcel and AQA-S); and over time (1999 and 2000).

These comparisons were realised primarily through the graphical and tabular representation of aggregate data, rather than through formal statistical testing. Too much emphasis upon statistical testing risks giving an impression of absolute objectivity which can be misleading when, as discussed earlier, assumptions necessary for valid comparison often do not hold. Likewise, formal statistical tests can misdirect attention towards statistically significant differences that are not necessarily important nor very meaningful. It is hoped that the graphical representations, in particular, have helped to emphasise the magnitude of any observed differences, thereby helping to signify their importance.
3.4.3.1 Patterns relating to type of review

In relation to the number of schools and pupils involved in review requests, with the exception of English, there was no clear tendency for more R1 requests to be submitted than R2 requests. However, for English, at both key stages, it was clear that more schools and more pupils per school were submitted for R2 review than for R1. Thus, concerns over mark scheme interpretation consistently outstripped concerns over clerical errors for English.

When review outcome changes were broken down into ‘up’ versus ‘down’, a clear difference between R1 and R2/R3 was evident. Pupils were hardly ever awarded lower levels following either R2 or R3 review. In this light, the threat in the Schools’ Guide that “the marks and the pupil’s level go either up or down” rings hollow. That so few pupils (3%) received lower levels following R3 review was especially surprising. It seems to suggest either that schools tend not to submit pupils near the lower thresholds of levels or that re-markers tend to shy away from awarding pupils lower levels. This is based on the assumption that a substantial percentage of R3 schools will have had inconsistent markers who would have been as likely to have made errors of lenience as of harshness. Rectified errors of lenience near to a lower threshold would lead to pupils’ levels being lowered. However, an alternative explanation might be that the vast majority of R3 reviews were requested because of consistently harsh marking, which might indeed result in far fewer pupils with levels lowered than levels raised.

A further difference between review types was the frequency with which outcomes resulted in changed levels. Across all subjects and both key stages, more R1 requests led to level changes than R2 requests. This is perhaps not surprising as clerical errors should, in principle, be relatively unambiguous. Indeed, perhaps we should be expecting higher percentages of R1 changes than the 73% for KS2 English or the 87% for KS2 science? Are there particular mistakes that schools are making that might be prevented through better instruction? Likewise for the failed R2 requests. An investigation into the reasons why review requests fail might prove illuminating.

3.4.3.2 Patterns relating to subjects

Differences in patterns of review requests between subjects were common. These were clearly related to the inevitable differences in marking reliability that result from different question formats and from more or less closed mark schemes.

Most obviously, more schools requested reviews in English than in science or maths; likewise more pupils were involved in such reviews and the mean numbers of pupils per
school's submission were higher for English. The only apparent contradiction to this trend was for R1 review at key stage 3, where there were fewer schools and fewer pupils than for maths and science. It is not obvious why this might have been observed (although it is probable that both the R1 and R2 figures for English at key stage 3 were somewhat deflated by the R3 review option).

Few consistent differences between maths and science were observed at key stage 2, although figures were higher for science than maths at key stage 3. Whether this might be a result of different approaches to marking between maths and science between key stages, or a result of differences in the nature of the subject content between key stages, or a result of other unrelated factors is not clear.

3.4.3.3 Patterns relating to key stage

Patterns in the number of schools and pupils involved in review requests were similar between key stage 2 and key stage 3. However, at least for English, the absolute numbers of pupils involved in reviews tended to be consistently higher at key stage 2 than at key stage 3. To some extent, this is likely to have been an artefact of the additional reviews in Reading and Writing at key stage 2. Yet it may also reflect the impact of performance tables at key stage 2.

Perhaps the most interesting difference between key stages related to the outcomes of review requests. Across all subjects at key stage 2, the percentage of R2 review requests that resulted in a change of level was 45%; at key stage 3, the percentage was 68%. This pattern of difference was consistent across subjects and there may have been a variety of factors that led to it. First, those responsible for checking KS3 pupils’ scripts (typically heads of subject departments) may have had a better grasp of the relevant subject matter content than those responsible for checking KS2 pupils’ scripts (typically headteachers). If so, we might expect them to have been more likely to identify mark scheme interpretation errors correctly. Second, it may be that key stage 3 teachers are generally more tolerant of mark scheme interpretation error and are, therefore, less likely to submit review requests in ‘marginal’ cases. Indeed, the pressure of performance tables at key stage 2 may have encouraged teachers to be less tolerant. To the extent that the latter is true, then the advent of performance tables at key stage 3 would be predicted to have a knock-on effect for the number and percentage of inappropriate review requests received.

3.4.3.4 Patterns relating to EMAs

As the number of review requests received by an EMA is likely to be a function of the quality of initial marking, it might be somewhat inappropriate to emphasise observed
differences during an evaluation of review services. In fact, though, few such differences were observed.

On the other hand, systematic differences in the outcomes of reviews between EMAs might be of much more relevance. As such, the percentages of pupils for whom levels were changed (versus not changed) following review were compared between EMAs for all subject-by-type-by-stage combinations (excluding KS3 English R3). Of the 12 comparisons, significant differences between EMAs were observed for six (all four English, one maths and one science). Across these six, no one EMA consistently produced the highest or lowest percentage of level changes, although Edexcel was responsible for the majority of the highest percentages.

If the quality of initial marking could be assumed to have been comparable between EMAs then this comparison of review outcomes might, at first glance, be taken to indicate the extent to which certain re-marking teams may have been more or less lenient between EMAs. On reflection, though, it is not clear how such reasoning could extend to the clerical checking process. As significant differences between EMAs were actually more prevalent for R1 reviews, it would probably be wise not to read too much into these between-EMA comparisons.

Differences between EMAs were also observed in relation to the dates by which review outcomes were returned. It should be noted that the return date data may not have been entirely consistent between (or even within) EMAs; therefore caution should be exercised when interpreting the data. However, from the best available evidence, it seems that differences between EMAs were apparent for R1, R2 and R3 reviews.

For R1 reviews, which can be turned around fairly rapidly, the most significant question was how many were returned before the end of term. The percentage of reviews despatched by the end of Friday the 28th July is probably the best index of this. Here, while AQA-N and Edexcel had despatched over 70% of reviews, AQA-S had despatched fewer than 30%. However, by the end of the following week, AQA-S had caught up, having despatched over 80%.

For R2 and R3, the comparisons became even more complicated and it was more apparent that the 'return date' did not necessarily correspond to the date of despatch. However, from the best possible evidence, it appeared that AQA-N was quicker that the other EMAs in turning around the majority of review requests.
3.4.3.5 Patterns relating to year

The final exploration of the review request and outcome data concerned trends over time. For these investigations, the QCA aggregate data were relied upon and an assumption was made that ostensibly comparable figures did represent the same thing each year. Particularly for key stage 2 English, there may be at least a small amount of doubt here.

With the exception of R2 reviews for English at both key stages, the number of schools that requested reviews and pupils for whom reviews were requested either remained fairly stable over the past few years or tended to change erratically. For R2 reviews in English, however, there was a more consistently increasing trend. Indeed, not only was there evidence of more schools and pupils over time, but there was also evidence that, on average, more pupils were being submitted by each school over time. Clearly, this trend ought to be monitored and may require further investigation.
Section 4 An investigation into schools’ views of the national curriculum review system (Study 3)

4.1 Introduction

To determine whether the review system was effective, fair and consistently operated it was essential to incorporate the perspective of schools. Studies 3 and 4 did so in slightly different ways. While Study 4 focused upon only those schools that had been through the review process during 2000, Study 3 adopted a broader approach and considered the views of a random sample of schools.

In one sense, contacting a random sample of schools might seem somewhat inappropriate; what could schools that had not necessarily utilised the review system tell us about its effectiveness? Yet, there are certain questions that it would be important to ask of a representative sample of schools (as a sample restricted to those that had requested reviews might be biased in favour of negative impressions). Moreover, there are certain other questions that can only be asked of a sample of schools which includes those that have not requested reviews; principally, questions that explore reasons for not requesting reviews.

The intention of Study 3 was to contact a large sample of schools in order than a genuine impression of the national perspective might be presented. As such it was decided to use a survey methodology.

4.2 Methodology

4.2.1 Sample selection

The first methodological issue to be resolved was how to construct the sample. In fact, it was decided to construct four samples: one for primary schools and three for secondary schools. The differential approach between key stages reflected the intention to target the person within each school who would be most likely to have direct responsibility for instigating review requests. Thus, it was decided to create one questionnaire for Headteachers of primary schools, covering English, maths and science, and three questionnaires for Heads of Departments of secondary schools, covering English, maths and science, respectively. The samples were to be selected such that no secondary school would receive more than one version of the questionnaire.

The population from which the samples were to be drawn was to include all primary and secondary schools in England with the exception of Special Schools, Pupil Referral Units and schools not participating in the national curriculum assessment scheme.
To ensure that the four samples were representative of all schools in that population, a stratification technique was employed. The samples were stratified by the following three factors:

1. **school performance** (from the 1998 primary and secondary performance tables). Schools were divided into 5 categories: from the lowest scoring 20% to the highest scoring 20% (based upon KS2 results or GCSE results).

2. **school type.** Primary schools were divided into 4 categories: Primary/Combined; Junior; Middle; and Independent. Secondary schools were divided into 7 categories: Middle deemed secondary; Secondary Modern; Comprehensive to 16; Comprehensive to 18; Grammar; Independent; and Other.

3. **region.** Schools were divided into 3 categories: North; Midlands; and South.

This ensured that the samples and the populations from which they were drawn were similar in terms of: the distribution of school types; the distribution of performance profiles; and the regional distribution.

The size of the sample of primary schools was 1,500. For secondary schools, the sample sizes were 400 for English, 300 for maths and 300 for science. The primary sample was larger than the secondary sample because the primary questionnaire was to be longer than the secondary ones, which might have led to a lower response rate. The fact that the secondary English sample was larger than the maths and science samples reflected the heightened concern with the marking of English.

### 4.2.2 Questionnaire development

The questionnaires were designed to address a number of issues which included: satisfaction with information on the review process; past experiences of requesting reviews; reasons for not requesting reviews; etc.. Although four versions were prepared, they all followed a basic template and shared as much as possible in common (so as to facilitate comparison between versions). Question response formats were designed to be as closed as possible (i.e., tick-boxes), to ensure that the survey was not too time-consuming for schools to complete. Unfortunately, the time-scale for developing the questionnaire was extremely tight and this meant that there was no time to pilot draft questions.

A copy of each of the four versions of the questionnaire can be found in Appendix 4.1.
4.2.3 Questionnaire distribution

Before the questionnaires were distributed, letters were sent to prospective LEAs (which included lists of prospective schools) to seek their approval. (This is standard NFER practice to ensure that undue demands are not placed on schools which face particular problems.) Questionnaires were then distributed on the 14th July, by which time most schools would have received their test results but before which few schools would have broken up. For both primary and secondary samples, the questionnaires were sent to schools with cover letters addressed to the Headteacher. For the secondary samples it was suggested that the Head of the relevant subject department would probably be best placed to complete the survey and Headteachers were encouraged to pass all documentation to them. For the primary sample it was suggested that the Headteacher might be best placed to complete the survey, although they were encouraged to consult with other teachers where, or if, necessary.

Reminder letters, which included a copy of the original questionnaire, were despatched to schools that had not responded to the original survey by 1st September, the beginning of the autumn term. The last batches of returned questionnaires were sent to the NFER Data Entry department by 21st September. Here they were double-entered, to ensure the fidelity of the data capture.

4.3 Results

4.3.1 Response rates

The response rates across all four samples were good, particularly bearing in mind the fact that the questionnaires were (unavoidably) despatched at the very end of the Summer term. Rates varied between 51% (KS3 maths) and 60% (KS2).

One LEA requested that none of their schools should be approached, citing the late timing of the survey as the reason. Similarly, there was a stronger than usual reaction from schools to the timing of the survey.

The full response-rate breakdown is presented in Appendix 4.2.

4.3.2 Reply-sample representativeness

The first task was to determine whether the sample that replied to each questionnaire was representative of the sample to which each questionnaire was sent (each of which had previously been constructed to be representative of the general population of regular, national curriculum schools in England).
In order to determine whether the four 'reply-samples' were representative of the four 'despatch-samples', each was broken down by the three stratification factors presented earlier and the relative proportions of respondents within the categories of each factor were examined (see Appendix 4.2).

Statistical tests were used to determine whether the proportions of respondents in each category of each factor for each reply-sample matched the proportions of respondents in each category of each factor for each despatch-sample. Only one of twelve 'Likelihood' tests (4 samples x 3 factors) gave any indication of significant differences. This was the breakdown of the KS3 English sample by the school performance factor. Even then, the differences were very small.

These results suggested that the four reply-samples were highly representative of the four despatch-samples in terms of the three stratification factors: school performance; school type; and region. It can therefore be inferred that, in terms of these three factors, the reply-samples were highly representative of the general population of regular, national curriculum schools in England.

In the following discussion, the term 'sample' will generally be used to refer to the sample of schools that responded to each questionnaire, rather than the sample of schools to which each questionnaire was sent: the reply-sample rather than the despatch-sample.

4.3.3 Responses to the survey questions

In general, the questionnaires functioned effectively and as intended. The vast majority of respondents answered the vast majority of the questions that they were invited to. The following discussion will highlight the main findings that emerged. A full breakdown of responses to individual questions can be found in Appendix 4.2.

4.3.3.1 Characteristics of the respondents

Before exploring the substantive results in detail, it is worth pausing to consider the nature of the sample of respondents in a little more depth using data that were collected from the questionnaires. This will help to contextualise the results by further characterising the kind of respondents from whom they were gathered.

4.3.3.1.1 EMA distribution (Q12)

An additional test of the representativeness of the reply-samples was whether the relative proportions of respondents from each EMA matched the national profile. This was not explicitly controlled for in the despatch-samples as the NFER had no prior knowledge of which schools were assigned to which EMAs. Question 12 asked schools to record the
EMA to which they were allocated and these data were used to evaluate the reply-samples.

Interestingly, it was apparent that a significant number of respondents did not know which EMA was responsible for the marking of their pupils' scripts. Over one-quarter of primary schools failed to respond to this question, as did between 8% (English) and 19% (science) of secondary schools. Unfortunately, this meant that an analysis of representativeness by EMA would be ambiguous because of the possibility that non-respondents would over- or under-represent particular EMAs.

However, when data from only those schools that responded to this question were inspected, there was broad agreement between the national picture and the characteristics of each reply-sample. For KS3 English and KS3 science the relative proportions were very similar to the national picture. The KS2 reply-sample tended to over-represent Edexcel slightly, at the expense of AQA-S. The KS3 maths reply-sample tended to over-represent AQA-N slightly, at the expense of Edexcel. Generally speaking, it appeared that the reply-samples were representative of the population of regular, national curriculum schools in England in terms of EMA.

4.3.3.1.2 History of requesting reviews (Q5, Q6a, Q6b)

Questions 5, 6a and 6b explored the extent to which respondents had used the review system in the past or during 2000.

At KS2, the proportions of respondents that ticked either the R1 or R2 box for 1998 were 9% (English), 6% (maths) and 8% (science). For 1999, the corresponding proportions were 14% (English), 8% (maths) and 7% (science).

At KS3, the proportions of respondents that ticked either the R1, R2 or R3 box for 1998 were 23% (English), 13% (maths) and 19% (science). For 1999, the corresponding proportions were 30% (English), 16% (maths) and 28% (science).

Generally speaking, more schools reported having requested reviews in 1999 than in 1998. Far more importantly, though, for all-but-one of the subjects at KS2 and KS3, the proportions arising from the questionnaires were within a few percent of the national figures for 1999. (The exception was KS3 science where there were around 10% more review requests in the reply-sample than in the national data for 1999.) This is good evidence that the reply-samples did not over-represent schools that had had negative experiences of national curriculum marking in the past (with the possible exception of KS3 science).
Finally, the proportions of KS2 respondents that had made, or that were considering making, review requests during 2000 were 18% (English), 7% (maths) and 8% (science). At KS3 the proportions were 44% (English), 20% (maths) and 34% (science). At KS2, these figures were similar to the national figures for 2000. At KS3, the proportion for maths was around 5% higher than the national proportion while, for KS3 English and science, the survey figures were around 10% higher than the national figures. Thus, while the KS2 sample was representative of the nation in terms of the proportion of review requests made during 2000, the KS3 samples appeared to have slightly over-represented schools that had made review requests.

4.3.3.1.3 Position in school (Q13)

While it was assumed that the majority of primary respondents would be Headteachers, it was decided to investigate exactly who filled in the secondary questionnaires. As expected, the vast majority of respondents were Heads of subject departments. The majority of those who were not Heads of subject departments still tended to be responsible for subjects at KS3.

4.3.3.1.4 Practical marking experience (Q11)

The vast majority of primary respondents ticked none of the boxes for Question 11. Only 10% said that they had any previous experience of external marking, mainly at KS2.

In contrast, only half of secondary respondents ticked none of the boxes. One-fifth of English and science respondents said that they had KS3 external marking experience and three-tenths of maths respondents said so. One-third of maths and science respondents said that they had GCSE marking experience and one-half of English respondents said so.

4.3.3.2 What do schools think of the quality of initial marking? (Q1, Q3a, Q3b, Q3c, Q4a, Q4b, Q4c, Q9a)

The number of review requests lodged in any year is a reflection of how schools perceive of the quality of initial marking. Yet, it is not a direct reflection. Schools may consider the quality of marking to be poor but not request any reviews; or they may consider the quality of marking good but still request a few reviews anyway. As such, it is essential to approach the issue directly. Question 1 did precisely this, asking how respondents would rate the quality of marking, in different subjects, during 2000 (on a five-point scale from 'Very poor', scored as 1, to 'Very good', scored as 5).

The average responses across subjects tended to cluster around the higher side of the middle 'Satisfactory' position (scored as 3). The average response was lowest for KS3 English (mean of 2.6) and highest for KS3 maths (mean of 3.9). More explicitly, no more
than 8% of respondents for maths or science at either key stage rated the quality of marking as ‘Poor’ or ‘Very poor’. For KS2 English the proportion was 13% while, for KS3 English, the proportion was 45%.

Question 9a posed a similar question, asking whether most scripts receive the marks that they deserve during the initial marking phase. Fewer than 8% of maths and science respondents either ‘Disagreed’ or ‘Strongly disagreed’. However, 16% of the KS2 English sample disagreed, as did 38% of the KS3 English sample.

Figure 1. Question 9a. “I believe that most [English] scripts receive the marks they deserve during the initial marking phase.”

Questions 3 and 4 attempted to quantify the extent of the perceived marking errors. Seventy three percent of KS3 English respondents believed that at least one of their pupils may have received too few marks during the initial marking phase; the average proportion of pupils believed to have been awarded too few marks was 26% and the average proportion of pupils believed to have been awarded too low a level was 17%. Conversely, only thirty one percent of KS3 English respondents believed that at least one of their pupils may have received too many marks during the initial marking phase; the average proportion of pupils believed to have been awarded too many marks was 8% and the average proportion of pupils believed to have been awarded too high a level was 9%.
As can be seen from Appendix 4.2, similar trends were observed in Questions 3 and 4 as were observed in Questions 1 and 9a. Marking was rated to be worst in KS3 English, then KS2 English, then the other subjects. The differences in the proportions of pupils awarded too few versus too many marks/levels were less pronounced for subjects other than KS3 English.

4.3.3.3 How do schools check the marking of test scripts? (Q2)

If a school is to lodge a review request then it must, at least, check the marking of scripts that it intends to submit. Yet, it is not recommended to schools exactly how they should check scripts; nor is it known how schools vary in their checking practices. Question 2 sought to explore this.

Interestingly, a small number of schools did not check the marking of scripts at all. This ranged from 2% for KS3 English respondents to 11% for KS3 maths respondents. Across all subjects, the majority of schools checked scripts by comparing between expected and observed pupil levels; this was the most frequently recorded check. Indeed, the most frequently recorded response-pattern was for schools to check only the comparison between expected and observed pupil levels (one-fifth of KS2, KS3 English and KS3 maths respondents, respectively).

The second most frequently recorded response-pattern was to check expected and observed levels, selected scripts for R1 errors and selected scripts for R2 errors (between 13% and 19% of schools across the four questionnaires). Three other response-patterns were similarly common; firstly, expected versus observed levels and selected scripts for R1 errors; secondly, expected versus observed levels, all scripts for R1 errors and selected scripts for R2 errors; thirdly, selected scripts for R1 errors and selected scripts for R2 errors.

4.3.3.4 What motivates schools to request reviews? (Q7, Q9b, Q9h)

One important objective of the survey was to explore reasons for requesting or not requesting reviews. Question 7 listed a variety of reasons and asked respondents to rate each in terms of the extent to which it would constitute a motivation to request an R2 review. Answers were recorded on a sliding scale from 1 (not at all significant) to 5 (very significant).

The first motivation was to ensure that all children scored as highly as possible. Across all subjects this was consistently rated highly, receiving average scores between 4.1 and 4.3. Indeed, this was the most highly rated of factors.
The second reason was to ensure that the school or department scored as highly as possible. Although slightly less of a motivation than pupil performance, this was still rated highly across all subjects, receiving average scores between 3.3 and 3.9. Only a minority of schools gave this an ‘insignificant’ rating of 1 or 2 (from 14% for KS3 science to 26% for KS3 English).

The third reason was to register a formal complaint against the system. Interestingly, schools tended not to consider this a significant motivation to request a review. Average responses ranged from 1.9 to 2.4 and the proportion of respondents that recorded an ‘insignificant’ rating of 1 or 2 ranged from 56% (KS3 English) to 72% (KS3 science).

**Figure 2.** Question 7b. How significant a motivation would be: ensuring that the school maximises its position in performance tables?

As for the second, the fourth – discrepancies between test and teacher assessment results – was deemed to be a considerable motivation to request a review. The proportion of respondents that recorded either 4 or 5 ranged from 43% (KS3 maths) to 73% (KS3 English).

Finally, the fifth reason, that appeared only in the KS2 questionnaire, referred to the fact that KS2 scores are input measures to value-added calculations. The implication of this is that achieving higher KS2 scores would improve a school’s value-added standing when comparing from baseline assessment to KS2 (which might, therefore, encourage schools...
to lodge review requests). With a mean rating of 3.5, respondents indicated that this tended towards a significant motivation.

Two further questions highlighted factors that might also be significant in motivating a school’s decision to request a review. The first, from 9b, concerned the extent to which respondents were prepared to tolerate a ‘small amount’ of marking error. Interestingly, there appeared to be less tolerance at KS2 than at KS3. At KS3, roughly one-third of respondents, across each of the three subjects, either ‘Disagreed’ or ‘Strongly disagreed’ with the statement that they were prepared to tolerate a small amount of marking error. Yet, the majority of KS3 respondents, across each of the three subjects, were in agreement with the statement. Conversely, at KS2, the majority of respondents — nearly three-fifths in fact — were in disagreement.

The second question, 9h, concerned the statement that a significant number of schools use the review process more to increase their results than to see justice done for their pupils. Once again, response-tendencies differed between KS2 and KS3, with a greater proportion of primary respondents agreeing with the statement. Between 22% (KS3 English) and 32% (KS2) of respondents across subjects either ‘Agreed’ or ‘Strongly agreed’ with the proposal. However, similar proportions of respondents disagreed with the proposal, between 22% (KS2) and 37% (KS3 English). Across all subjects, most respondents were undecided.

4.3.3.5 What discourages schools from requesting reviews? (Q8)

Question 8 was, essentially, the converse of Question 7. It listed a variety of factors and asked respondents to rate each in terms of the extent to which it would constitute a discouragement from requesting an R2 review. Once again, answers were recorded on a sliding scale from 1 (not at all significant) to 5 (very significant).

The first factor was the complexity of the R2 review procedure. Across all subjects, responses tended to average to the half-way mark between ‘not at all significant’ and ‘very significant’. However, this concealed a considerable dispersion of views, with large numbers considering the complexity of the procedure very significant (between 9% and 25%) and large numbers considering it not at all significant (between 17% and 35%). Most concern was evident from KS3 English respondents and least concern from KS3 maths respondents.

The second factor was the time taken to check scripts and to prepare an R2 review case. For this question, responses tended to average just above the half-way mark between ‘not at all significant’ and ‘very significant’. Once again, there was a considerable dispersion.
of views: while around one-half of respondents to each questionnaire gave this factor a rating of 4 or 5, between one-fifth and one-third responded with 1 or 2. Again, most concern was expressed by KS3 English respondents.

**Figure 3.** Question 8a. How significant a discouragement would be: the complexity of the R2 review procedure?

The third factor was the fact that a pupil’s level could, in principle, go down. Of all the factors in Question 8, this received the lowest average ratings, which ranged from 2.4 to 2.6 across subjects. Between 50% and 55% of respondents gave this factor a rating of either 1 or 2.

Responses to the fourth factor, having to pay the deposit of £5 per pupil for R2 reviews, were very consistent across the four questionnaires. For each questionnaire, a similar proportion of respondents ticked each response category, representing a very clear dispersion of views and a lack of agreement between respondents.

In contrast, there was more consensus over the sixth factor, which was if the deposit happened to be set at £20 per pupil for R2 reviews. Across questionnaires, between 48% and 59% of respondents said that this would be a ‘very significant’ discouragement. However, between 8% and 13% still responded that this would be ‘not at all significant’.
The final factor concerned the recognition that scores are input measures to value-added calculations. This question was only included in the secondary questionnaires. The assumption was that schools might be discouraged from requesting reviews because higher value-added from KS3 to GCSE is facilitated by keeping KS3 scores low. (Of course, in principle, schools could request reviews to bring pupils levels down; in practice, though, this tends not to happen in relation to R2s.) Once again, there was both a similar pattern of responses across questionnaires and considerable dispersion of responses within each questionnaire. Thus, responses tended to average to the middle rating between the extremes of 'not at all significant' and 'very significant'. Around one-fifth of respondents for each questionnaire ticked the least significant extreme and around one-fifth ticked the most significant extreme.

4.3.3.6 Who do schools involve in the decision to request a review? (6c)

An important aspect of understanding the nature of review requests is understanding who is involved in the decision-making process. Question 6c explored this issue. Generally speaking, the people most involved in decision-making were Headteachers, Senior Management Teams, Heads of Departments (KS3) and Teachers. Governors and LEA officers were rarely involved. Parents were involved by a small number of schools; pupils were involved in up to one-fifth of schools (ranging between 3% for KS2 maths and 20% for KS3 science).

There were some notable differences between key stages and between subjects. For example, while Headteachers were involved in around nine-tenths of review decisions across subjects at KS2, their involvement tended to be much less at KS3 and tended to differ between subjects (from 4% for KS3 maths, to 30% at KS3 science, to 61% at KS3 English). At KS3, the individuals most involved by schools were Heads of subject departments, who were involved to the same extent as Heads of primary schools. Senior Management Teams were involved to a roughly similar extent across subjects and key stages (from 21% of schools for KS3 maths to 47% for KS2 English and science). Teachers were also involved to a roughly similar extent across subjects and key stages (from 61% for KS3 maths to 82% for KS3 English).

At KS2, across all subjects, the most common response-pattern was for decisions to involve Headteachers and Teachers only. The second most common response-pattern, again across all subjects, was for decisions to be taken by Headteachers, Senior Management Teams and Teachers only. At KS3, the most common response-pattern for maths and science was for decisions to involve Heads of subject departments and Teachers only; for English, Headteachers were also involved. Generally speaking there
tended to be more people involved in decisions for English at KS3 than for maths or science.

4.3.3.7 Do schools understand the review system? (Q9e, Q10a, Q10b, Q10c)

An important principle underlying national curriculum assessment is that it should be transparent, which might be interpreted as meaning, above all, that stakeholders should understand it. Questions 9e and 10 explored the extent to which respondents understood the national curriculum review system.

Question 9e presented the statement 'I understand the stages involved in making a review request' and asked respondents to rate it using a five-point scale from 'Strongly disagree' (scored as 1) to 'Strongly agree' (scored as 5), with a middle point of 'Undecided' (scored as 3). The reaction was encouraging, with two-thirds of KS3 maths and science respondents ticking the 'Agree' or 'Strongly agree' boxes, and with four-fifths of KS2 and KS3 English respondents ticking them. No more than 7% of respondents in any of the subjects ticked the 'Strongly disagree' box.

Figure 4. Question 10b. How clear was it to you how to complete the review request?

![Bar chart showing percentage of valid responses for KS2 and KS3 subjects.]

- Clear (VC/C)
- Satisfactory (S)
- Unclear (UV/U)
Question 10a asked whether respondents were aware of having received information from their EMA explaining procedures for requesting a review. Over two-thirds of KS2 and KS3 English respondents said that they had, although only just over one-half of KS3 maths and science respondents said so.

Questions 10b and 10c were restricted to only those respondents who said 'yes' to 10a. They were first asked how clear it was to them how to complete a review request. No more than 8% of respondents in any of the subjects ticked the 'Unclear' or 'Very Unclear' boxes. Indeed, between one-half and two-thirds of respondents ticked the 'Clear' or 'Very clear' boxes.

When asked how clear it was to them how the External Marking Agencies would undertake the review, the reaction was still positive, although slightly less so. Between two-fifths and three-fifths of respondents responded with 'Clear' or 'Very clear'. KS3 English respondents were the least positive, with one-fifth responding with 'Unclear' or 'Very unclear'.

4.3.3.8 What do schools think of the review system? (Q9c, Q9d, Q9f, Q9g, Q9l)

Five of the statements within Question 9 concerned the effectiveness and fairness of the review system. Once again, the response categories ranged from 'Strongly disagree' (1) to 'Strongly agree' (5), with a middle point of 'Undecided' (3).

Question 9c presented the statement 'I believe that most scripts that undergo an R1 review end up with the correct level'. For all subjects, more than half of respondents agreed. Yet, while no more than 8% of respondents in any of the subjects disagreed, a significant number in each of the subjects were undecided (between 22% for KS3 science and 39% for KS3 English). When asked the same question regarding R2 reviews (9d), fewer respondents agreed and more were undecided. Only for KS3 maths was the vast majority of respondents in agreement (72%).

Question 9f presented the statement 'I believe that the review system is generally fair' and 9g presented the similar statement 'I believe that the review system is sufficiently impartial and independent'. Perhaps not surprisingly, responses to these two questions were almost identical within subjects. The least positive response came from KS3 English where only two-fifths of respondents were in agreement with the statements and two-fifths were undecided. Responses for KS3 maths and science and KS2 were similar to each other, with around two-thirds of respondents in agreement and around one-third undecided (and only a very small minority in disagreement).
Figures 5 and 6. Question 9c and 9d. "I believe that most scripts that undergo an R1 [R2] review end up with the correct level."
Finally, when presented with the Question 9l statement ‘I believe in the right to request a marking review’, no more than 1% of respondents in any of the subjects disagreed. Between 62% and 87% of respondents ‘Strongly agreed’ while virtually all of the remainder ‘Agreed’.

4.3.3.9 What do schools think of the EMAs? (Q9i, Q9j, Q9k)

Three questions, 9i, 9j and 9k, presented statements of satisfaction with the External Marking Agencies. The first stated ‘My External Marking Agency is approachable’; the second stated that it was ‘trustworthy’; and the third that it was ‘efficient’. Once again, the patterns of responses were similar between the three questions and, generally speaking, the responses were not negative. The results presented below describe only the overall patterns; differences between EMAs are described later.

For the KS2 questionnaire, all three questions received around half of all responses in the ‘Undecided’ category and half of all responses in the ‘Agree’ or ‘Strongly Agree’ category. No more than 5% of responses were negative for any of the three questions.

Figure 7. Question 9k. “My External Marking Agency is efficient.”

Similarly, for KS3 maths and science, no more than 8% of responses were negative for any of the three questions. Response-patterns were almost identical for the KS3 maths and science samples in relation to the ‘trustworthy’ and ‘efficient’ statements; around
two-thirds of respondents were in agreement and around one-third were undecided. For the ‘approachable’ statement, respondents tended in the opposite direction, with just over two-fifths of KS3 maths and science respondents in agreement and just over one-half of respondents undecided.

As for the KS2 respondents, results for the KS3 English respondents were similar across all three questions. However, results from this sample were the most negative. Only one-third of respondents were in agreement with the three statements, while one-half were undecided. Across questions, between 10% (‘trustworthy’) and 27% (‘efficient’) of KS3 English respondents disagreed with the statements.

4.3.3.10 Additional issues that arose (Q13 or Q14)

At the end of the four questionnaires, respondents were encouraged to record any further thoughts that they were keen to express. The intention of this section was primarily to ensure that the questionnaire had tapped all of the issues that schools considered to be important (particularly as there had been no formal pilot). Additional comments were provided by approximately one-fifth of KS2, KS3 maths and KS3 science respondents; nearly two-fifths of KS3 English respondents provided comments. With relatively few exceptions, the comments were negative, focusing on perceived shortcomings rather than upon strengths of the system or upon scope for improvement.

At KS3, the most frequently raised issue was the perceived lack of time available for schools to check scripts and prepare review requests. Similarly, but less frequently, it was commented that the amount of work required to produce review requests was prohibitive. For KS3 English, a number of respondents explained that they had lost faith in the external marking system; having found significant errors in their pupils’ scripts year-on-year, they felt that the markers were simply not good enough. No further comments were raised by more than a few KS3 respondents.

At KS2, by far the most frequently raised issue was concern with the quality of marking of KS2 English and the perceived subjectivity of the mark scheme. Concern was most evident with respect to Writing and Handwriting. The second most frequent comment was the lack of time available for checking scripts and preparing requests, mirroring concern at KS3. Amongst the other issues raised by more than a few KS2 respondents were: that scripts should be checked better before being returned to schools; that the deposit was wrong in principle; that there should be an option for a full re-mark of scripts; and that results are needed by schools earlier.
A few quotations are presented below to give a feeling for the kind of comments received. As comments came from only a minority of respondents, they should not be taken to be representative of the views of all schools.

"It happens too late. Once a pupil grade is published for pupils/parents any damage caused by a low incorrect grade would already be done, any correction later does not compensate for pupil disappointment." KS3 science

"Having found a 12 mark discrepancy with one candidate, I have, to some extent, lost faith in the system. I object to having to re-mark the scripts in order to obtain a review and I strongly object to having to pay to rectify the mistakes of employed markers. I am concerned that only errors which change levels are reviewed. Markers should not make any significant errors – appropriate checking procedures should be in place." KS3 maths

"We have requested reviews in four of the past five years. The outcome has been unsatisfactory. Given the meaninglessness of KS3 English results and the limits on time and energy at this stage of the term we simply haven’t the appetite to request reviews this year.” KS3 English

"The time required of teachers to check scripts is unfair. The cost for the school is too great. The whole procedure really lowers morale. We feel quite angry about it.” KS3 English

"Marking of English writing (including handwriting) appears less objective than other subjects. I would feel happier if ALL writing scripts were marked twice – by a different marker.” KS2

"There is NO time at the end of the academic year to analyse any of the marking of the scripts unless there is a notable discrepancy between TA and SAT results. We may be doing our children/school a disservice by not checking the marking but, unfortunately, so be it. We need TIME!!!” KS2

4.3.4 Comparing responses between the External Marking Agencies

In addition to the overall breakdown of responses, which highlighted certain differences between subjects, it was decided to explore responses to a number of questions by EMA; that is, to explore whether responses to certain questions differed according to which EMA was responsible for marking schools’ scripts and processing review requests. The questions that seemed most appropriate to consider were 1, 6a and 9 (which focused upon quality of initial marking, prevalence of review requests and views of the system,
respectively). Before discussing the analyses further it is important to stress some very important caveats.

First, schools are not assigned to EMAs randomly. Indeed, they are assigned on an explicitly non-random basis: geographical location. This means that we have no way of being sure whether, say, Edexcel respondents are of a similar kind to AQA-N respondents. If schools from the north of England tend to be more or less vociferous (for example) than schools from the south-east then questionnaire responses might identify differences apparently attributable to the EMAs that are actually attributable to the populations of respondents.

Second, the kind of statistical tests used for comparing between EMAs should only be viewed as indicative rather than as definitive. On the one hand, the ANOVA tests used to explore responses to Questions 1 and 9 are ideally employed with data that demonstrate qualities that cannot be expected of data generated from simple questionnaire responses. On the other hand, while the tests can determine whether or not apparent differences are statistically significant, they cannot indicate whether they are meaningfully different. For example, there might be a significant difference between how trustworthy AQA-N and AQA-S respondents perceived their EMA to be; yet, if both were perceived to be extremely untrustworthy, while Edexcel was perceived to be extremely trustworthy then we might not consider the difference between AQA-N and AQA-S worthy of mention or of further exploration. Finally, of course, as with any statistical analysis, apparently significant differences might still occasionally result from ‘chance’ variation alone. (In the present report, significance is defined in a conventional sense; if the probability of an observed difference having happened ‘by chance’ is calculated to be less than 5% then it is taken to be significant.)

4.3.4.1 The quality of initial marking and intention to request a review during 2000? (Q1, Q9a, Q6a)

Question 9a explored the quality of initial marking by asking respondents to express their agreement or disagreement with the statement that most scripts ‘receive the marks that they deserve during initial marking’. Before conducting a statistical analysis of responses it was decided to omit respondents who were ‘Undecided’ so that the new scale, from ‘Strongly disagree’ (1) to ‘Strongly agree’ (4), represented an unambiguous rank order of strength of feeling. This meant that Question 9 analyses would compare levels of agreement between EMAs (using one-way ANOVA tests) only for those respondents who expressed an opinion.
Figures 8 and 9.

Question 1. How would you describe the general quality of marking of key stage 2 test scripts during 2000?
The only subjects for which significant differences between EMAs were observed were for KS3 English and KS3 science. For KS3 English, while the mean ratings for AQA-S (2.28) and Edexcel (2.19) were not significantly different, the mean rating for AQA-N (2.69) was significantly higher than both the others. For KS3 science, while the mean ratings for AQA-N (3.21) and AQA-S (3.14) were not significantly different, the mean rating for Edexcel (2.87) was significantly lower than both.

Question 1 asked a similar question; respondents were required to record their impression of the quality of initial marking during 2000, using a five-point scale from ‘Very poor’ (1), through ‘Satisfactory’ (3), to ‘Very good’ (5). This continuous five-point scale was used as the basis for slightly different analyses (two-way ANOVA tests). At KS3, the results of these analyses were very similar to those presented above for 9a. For both KS3 English and KS3 science, while the mean ratings for AQA-S and Edexcel were not significantly different, the mean rating for AQA-N was significantly higher than both. At KS2, there were significant differences between EMAs that were apparent across English, maths and science: AQA-N received significantly higher average ratings than Edexcel, although the average ratings for AQA-S were significantly different from neither.

A further, although indirect, index of the quality of marking is the extent to which EMAs differ in the number of review requests received. Question 6a asked whether respondents had, or were intending to, lodge review requests during 2000. Across all subjects at both key stages, there were no significant differences between EMAs in the relative proportions of respondents that ticked the ‘yes’, ‘no’ or ‘maybe’ boxes.

4.3.4.2 What motivates schools to request reviews? (Q9b, Q9h)

Question 9b asked respondents to express their agreement or disagreement with the statement ‘I am prepared to tolerate a small amount of marking error’. Once again, as for all Question 9 analyses in Section 5.3.4, responses were re-coded from a 1-to-5 scale to a 1-to-4 scale, omitting the ‘Undecided’ response.

Although not immediately obvious why, significant differences between EMAs were observed for the KS2 questionnaire and the KS3 science questionnaire. For KS2, the mean Edexcel rating (2.05) was significantly lower than the mean AQA-N rating (2.25), while the mean AQA-S rating (2.09) which lay intermediate was significantly different from neither. For KS3 science, exactly the same pattern held. The mean Edexcel rating (2.39) was significantly lower than the mean AQA-N rating (2.88), while the mean AQA-S rating (2.55) was significantly different from neither.
Question 9h asked whether respondents thought that a significant number of schools requested reviews more to increase their results than to see justice done for their children. No significant differences in responses were observed between EMAs.

Figure 10. Question 9b. "I am prepared to tolerate a small amount of marking error."

4.3.4.3 Do schools understand the review system? (Q9e)

When schools were asked whether they understood the stages involved in making a review request there were no significant differences in responses between EMAs for any of the questionnaire versions.

4.3.4.4 What do schools think of the review system? (Q9c, Q9d, Q9f, Q9g, Q9l)

Questions 9c and 9d asked respondents to rate their agreement with the statement that most scripts which undergo a review end up with the correct level, for R1 and R2 reviews respectively. For both questions, significant differences between EMAs were observed for only one subject, KS3 maths. For R1 reviews, the mean rating for AQA-N (3.41) was significantly higher than the mean ratings for AQA-S (3.16) and Edexcel (3.10) which were not significantly different. Exactly the same pattern held for R2 reviews where the mean rating for AQA-N (3.43) was significantly higher than the mean ratings for AQA-S (3.13) and Edexcel (3.11) which were not significantly different.
For the remaining three questions, 9f, 9g and 9l – which concerned, fairness, independence and the right to request a review – no significant differences in mean rating were observed between EMAs for any of the subjects at either key stage.

4.3.4.5 What do schools think of the EMAs? (Q9i, Q9j, Q9k)

Questions 9i, 9j and 9k were most explicitly about the EMAs. The first asked respondents to rate the extent to which they believed their EMA to be ‘approachable’, the second ‘trustworthy’ and the third ‘efficient’.

For the KS3 English questionnaire, significant differences were observed between EMAs for all three questions. When asked if their EMA was approachable, the mean rating for AQA-N (3.00) was significantly higher than the mean rating for AQA-S (2.53) with the intermediate rating for Edexcel (2.73) significantly different from neither. When asked if their EMA was trustworthy, a similar pattern was observed, with the mean rating for AQA-N (3.10) being significantly higher than the mean rating for AQA-S (2.58). This time, however, Edexcel’s intermediate mean rating (2.75) was significantly lower than AQA-N but not significantly different from AQA-S. The pattern was repeated when respondents were asked if their EMA was efficient. The mean rating for AQA-N (2.83) was significantly different from the mean rating for AQA-S (2.20) and Edexcel (2.17), although the latter two were not significantly different.

Figure 11. Question 9k. “My External Marking Agency is efficient.”
For the KS2 questionnaire, significant differences between EMAs were observed for 9j and 9k. When asked whether their EMA was trustworthy, the mean rating for AQA-N (3.11) was significantly higher than the mean ratings for both AQA-S (2.98) and Edexcel (2.97). When asked whether their EMA was efficient, the mean rating for AQA-N (3.07) was significantly higher than the mean rating for Edexcel (2.86) although the intermediate mean rating of AQA-S (2.97) was significantly different from neither.

Finally, for the KS3 science questionnaire, significant differences were only observed when respondents were asked to rate the efficiency of their EMA. While the mean ratings for AQA-N (3.17) and AQA-S (3.15) were not significantly different from each other, both were significantly different from Edexcel (2.75).

4.4 Discussion

It was pleasing to note both the good response rates and the representativeness of the samples that replied to the questionnaires. It is fair to conclude that the views expressed in the four questionnaires were representative of all regular, national curriculum schools in England at the end of the Summer 2000 term. The caveat to this conclusion was the possibility that the KS3 samples may have slightly over-represented those schools that had had more negative experiences of marking during 2000.

The presentation of results had two main components, an overall analysis for all questions which considered differences between subjects and a more focused analysis for specific questions which considered differences between EMAs. It should be re-emphasised that any differences between subjects or between EMAs should be treated with caution, primarily because (for obvious reasons) responses were not assigned to these groups at random. As such, it would always be possible that observed differences in responses between groups could be explained by characteristics that were irrelevant to the questions at hand. For example, while it was assumed that the majority of KS2 questionnaires were completed by Headteachers, the majority of KS3 questionnaires were completed by Heads of subject departments. Thus, apparent differences between KS2 and KS3 might actually be attributable to general differences between Headteachers and Heads of subject departments. Likewise, apparent differences between EMAs might actually be attributable to general differences between people in different regions.

4.4.1 What do schools think of the quality of initial marking?

Generally speaking, respondents were happy with the quality of marking in maths and science at both key stages. Across questionnaires, no fewer than 92% of respondents believed that the quality of marking in maths and science was at least satisfactory.
However, for KS2 English this figure dropped to 87%, while for KS3 English it dropped to only 55%. Interestingly, these findings agree closely with those that emerged from a telephone survey of KS2 and KS3 teachers – conducted for SMR on behalf of QCA during May and June 2000 – which concerned the marking of the 1999 tests (SMR, 2000a,b). Clearly, then, there are problems with the perception of the quality of English marking amongst the teaching profession and these are most acute for KS3 English.

When responses to Question 1 were broken down by EMA, some interesting results emerged. While no differences were observed for maths at KS3, it appeared that AQA-N respondents were more happy with the quality of initial marking for KS3 English and science than both Edexcel and AQA-S respondents. Similar findings emerged from analyses of responses to Question 9a. The Question 1 analyses also revealed a similar trend in responses between EMAs at KS2, with AQA-N having received a higher average rating across subjects than Edexcel. It should be noted that the scale of differences between EMA ratings for KS2 were considerably smaller than those for KS3 English and science (where the average difference amounted to one-half of a point on the rating scale). Particularly for KS3 English and science, it may be worth considering whether these findings might be attributable to differences in procedural practice between EMAs.

When the nature and extent of marking errors were probed in more depth a number of interesting findings emerged. The first was that across all three subjects, at both key stages, a considerable proportion of respondents believed that at least one of their pupils had been awarded too many marks. This ranged from 3% for KS2 maths to 20% for KS3 English (not including the ‘maybe’ responses). Indeed, of these respondents, the average proportion of the year group believed to have been awarded too high a level as a result of this ranged from 2% for KS3 maths and science to 10% for KS2 English and science. In some senses this is a somewhat trivial finding – random marking errors would be as likely to result in pupils obtaining higher levels than they deserved as lower levels. In another sense, though, it addresses an issue of fundamental significance for an evaluation of the review system. When combined with the fact that requests for R2 reviews to lower pupils’ levels are very rare (see Study 2), it demonstrates that schools do not seek downgrades even when they are well aware that marking errors have led to inaccurate results.

The second interesting finding was that, across all three subjects at each key stage, the proportion of respondents who believed that marking errors had resulted in too low a mark (for at least one pupil) was consistently higher than the proportion of respondents who believed that marking errors had resulted in too high a mark (for at least one pupil).
Indeed, the difference in proportions, across subjects and key stages, was generally of the order of one-half.

One interpretation of this might be that, across all three subjects at each key stage, markers are genuinely more harsh than lenient in the application of mark schemes. An alternative explanation is that teachers may be biased towards seeing the best in their pupils’ work and are more likely to identify negative injustice than positive. If true, this identifies a significant barrier to the professional perception and understanding of assessment: one that is likely to result in teachers viewing the national curriculum tests as loaded against their pupils. Yet another explanation might be framed in terms of the psychological construct of cognitive dissonance. That is, respondents’ answers to the question about ‘too many’ marks may reflect their earlier answers concerning ‘too few’ marks and the fact that teachers generally do not request reviews for downgrades: in order to minimise any feelings of guilt about not requesting downgrades, teachers may alter their beliefs about the scale of the problem of pupils being awarded too many marks. Once again, this would identify a significant barrier to the professional perception and understanding of assessment.

Interestingly, other reports on the national curriculum testing process (including auditor reports and reports on marking trials) have identified similar phenomena. It may be that underlying the apparent bias is an interaction between teachers’ personal knowledge of individual pupils and their desire to see them achieve at least as well as they deserve to. For example, if a teacher was confident that a particular pupil had been working at level 5 prior to testing, then she might be more disposed to search for marking error in the face of a level 4 test result than in the face of a level 6. Indeed, she might even be more disposed to ‘see’ marking error where others would not.

4.4.2 How do schools check the marking of test scripts?

Marking errors cannot be put right unless they are identified. The fact that up to 11% of schools did not check their national curriculum test results is therefore worrying; it suggests that a significant number of pupils would have been disadvantaged through a lack of action on behalf of their school. Of course, this is not necessarily to be taken as a criticism of these schools; as a number of respondents mentioned, in the free-response section, the amount of time available for checking scripts was prohibitive.

However, the vast majority of schools did, at least, check the plausibility of levels awarded to their pupils. Indeed, around two-thirds of schools checked the marking of scripts for at least some of their pupils, either for R1 clerical errors or for R2 mark scheme
application errors. Yet, only the minority of schools checked all of their pupils’ scripts for R1 errors and even fewer checked all of their pupils’ scripts for R2 errors.

4.4.3 What motivates schools to request reviews?

An important role for the survey was to explore schools’ attitudes towards the review system and the reasons why they decide, or do not decide, to request reviews. Most of the factors that were identified as motivators were rated as such by schools. These included: ensuring that pupils, school and department score as highly as possible on the tests; wanting to deal with discrepancies between test and teacher assessment results; and, for primary schools, the potential impact of test scores on value-added calculations. Wanting to register a formal complaint was seen as less of a motivation. These findings seem to indicate the pressure upon teachers to maximise results for pupils and the school alike.

Whether the need to maximise results has negatively impacted upon the use of the review system is less clear from the survey. While a considerable number of respondents felt strongly that some schools were ‘playing the system’, a similar number felt strongly that this was not the case. Most, though, were undecided.

Presumably, review requests reflect an unwillingness to tolerate perceived marking errors in pupils’ scripts. As such, a relevant question might be the extent to which teachers are prepared to tolerate a small amount of marking error. That there appeared to be less tolerance at KS2 was interesting and, perhaps, significant in relation to the proportion of review requests rejected at KS2 (which was higher than at KS3). Whether the difference in responses between KS2 and KS3 bears any relation to the finding that primary respondents typically did not have external marking experience, while secondary respondents typically did, is worth considering. Perhaps increased tolerance of marking error is a consequence of having had direct experience of marking in the past.

4.4.4 What discourages schools from requesting reviews?

When asked about factors that discourage schools from making review requests, a characteristic finding was the dispersion of responses across the five response-categories from ‘Strongly agree’ to ‘Strongly disagree’. Clearly many of the factors that were highly significant for a large number of schools were not at all significant for a large number of others. This dispersion of views was typically replicated across subjects, although KS3 English responses were somewhat more polarised as they tended to see the factors as most problematic. These factors included: having to pay a deposit of £5 per pupil; the complexity of the review procedure; the time taken to prepare review cases; and, for secondary schools, the potential impact of test scores on value-added calculations. One
way of interpreting these response patterns is that only some schools saw the factors as de-motivating hurdles while others felt that they were not obstacles at all. Another is that all schools recognised the factors as de-motivating hurdles, but that only some schools were actually put off from requesting reviews by them.

It was interesting to note how the fact that a pupil’s level could, in principle, go down was rated the lowest of all in terms of discouraging the decision to request a review. To some extent, this may well reflect the fact that only pupils whose levels are likely to be increased are generally submitted for review – the worst that is likely to happen is that their level stays the same.

One factor that was generally perceived as a potential discouragement to lodging review requests was if the R2 deposit were set at £20 per pupil, rather than £5.

4.4.5 Who do schools involve in the decision to request a review?

Formally, the submission of a review request comes from a school’s Headteacher. Indeed, requests are returned if the Headteacher’s signature is not on the relevant form. Yet Heads were not universally reported to have been involved in the decision to request reviews (for those schools that had decided to do so during 2000). Between 9% (KS2E) and 96% (KS3M) of respondents who had requested reviews during 2000 claimed that the Headteacher was not involved. At KS3, the most frequently reported decision-makers were Heads of subject departments. This finding might have implications for whom EMAs should be dealing with in relation to marking reviews. Indeed, it might have implications for whom important information is disseminated to (i.e., the Schools’ Guide).

Despite rumours of LEA officers encouraging schools to request reviews where possible – to increase the LEA performance profile – it was only infrequently reported that LEA officers had any input into the decision. However, while little if any LEA input was recognised at KS3, there was slightly more recognition at KS2 (where performance tables are published). This might suggest a small grain of truth in the rumour.

The involvement of pupils in the decision-making process appeared to be slightly higher at KS3 than at KS2. It is not clear why this might be the case (one possible suggestion is a higher incidence of pupils checking through scripts against mark schemes during class time – although this is largely speculation, there was some supporting evidence from Study 4). Finally, it was interesting to note that, at KS2, up to 8% of respondents mentioned an input from parents. The nature and extent of this involvement might be worth exploring further.
4.4.6 Do schools understand the review system?

It was reassuring to note that survey respondents generally felt that they understood the stages involved in making a review request. Furthermore, of those who had requested (or who were intending to request) reviews during 2000, the majority felt that the instructions for doing so were clear (and responses were similar across subjects and key stages). SMSR asked a similar, but more general, question about the clarity of the Schools’ Guide for 2000 (SMSR, 2000a and b). They noted a similar level of satisfaction with the clarity of the document, although slightly higher levels of satisfaction amongst KS3 respondents.

The majority of survey respondents also felt that they understood how their EMA would undertake their review request. While these findings are encouraging it should not be forgotten that a significant number of respondents still suggested that the guidance was unclear or that they did not understand how the review would be undertaken. If schools are to submit reviews correctly and are to consider the review process transparent then steps must be taken to further improve the quality of guidance and explanation.

4.4.7 What do schools think of the review system?

One of the most direct questions on the survey was whether or not respondents agreed with the statement ‘I believe that most scripts that undergo an R1 review end up with the correct level’. While few respondents disagreed, between one-fifth (KS3S) and two-fifths (KS3E) of respondents were ‘Undecided’. When the same question was asked in relation to R2 reviews, similar results were observed, although now the proportions ‘Undecided’ ranged between one-third (KS3M) and one-half (KS2E). This is not a particularly reassuring outcome. It seems to suggest either that respondents did not feel they had enough information to decide, or that the information that they were in possession of was indeterminate. Either way, there is a clear need for some kind of public-relations work in this area. In the same way that marking reliability studies are essential for upholding confidence in the rigour of initial marking, there is a strong case for the same kind of study applied to review marking.

Interestingly, when responses to these questions were compared between EMAs, the only significant differences were observed (on both questions) for mathematics, where AQA-N received higher average ratings than AQA-S and Edexcel. Whether these might meaningfully be interpreted in terms of procedural differences between EMAs is not clear.

When asked whether the review system was generally fair and sufficiently independent and impartial the response, once again, was cautiously positive. While the majority of respondents were in agreement, around one-third of respondents, across questionnaires,
were ‘Undecided’. This is further evidence of the need for more attention to be paid to the professional perception of the review system.

Finally, while there was some evidence of uncertainty that the review system was functioning effectively at present, there was almost unanimous agreement that schools should have the right to access such a mechanism.

4.4.8 What do schools think of the EMAs?

The first point to highlight in this respect is the considerable number of respondents who appeared to be unaware of which EMA was responsible for the marking of their scripts. While not entirely clear why, it is possible that this may reflect some confusion between AQA-N and AQA-S.

Having mentioned this, we can turn to the substantive issue of responses to the three questions that asked whether respondents agreed that their EMA was ‘approachable’, ‘trustworthy’ and ‘efficient’. Importantly, as the sample was not restricted to schools that had just been through the review process, responses should be taken to indicate general impressions of the agencies rather than impressions of their review services, per se.

Although their perceptions of the EMAs were generally not negative, a substantial number of schools were ‘Undecided’ in response to all three questions. The proportions of respondents who were in agreement with the statements often did not constitute the majority. Once again, this highlights a problem with the professional perception of the external marking system that needs to be addressed. Once again, the most negative impressions were expressed by KS3 English respondents. And, once again, whenever significant differences were observed between EMAs, they were always in favour of AQA-N.
Section 5 A ‘customer satisfaction’ perspective on national curriculum review services during 2000 (Study 4)

5.1 Introduction

While the intention behind Study 3 was to explore how schools in general perceive the review system, Study 4 was intended to focus more specifically upon only those schools that had requested reviews during 2000. In many ways, then, it was to be an investigation into customer satisfaction.

The major methodological decision was how to conduct such an exercise. As it was felt necessary both to draw conclusions from a broad representative sample and to obtain rich qualitative information, it was decided to adopt more than one technique. Thus, a large-scale postal survey followed by a small-scale telephone survey was planned.

It was also felt that the most interesting information would come from those schools that had submitted either R2 or R3 reviews – individual or group re-marks. While R1 reviews – clerical checks – are a central feature of the system, they involve far less ambiguity and, therefore, are far more straightforward to resolve and are far less likely to result in customer dissatisfaction. In this sense, then, Study 4 was intentionally biased; it was to focus attention on the potentially more problematic aspect of the review system.

It was intended originally to produce only two questionnaires: one for all R2 reviews (across subjects and key stages) and one for R3 reviews. However, it became apparent that this would introduce an unnecessary element of obfuscation, as numerous subtle differences between key stages and subjects would have had to have been accommodated. Therefore, it was decided to produce seven questionnaires: one R2 version for each subject at both keys stages and one R3 version for English at key stage 3. Despite this, it was intended that the questions on each questionnaire should be as similar as possible to facilitate comparison of responses across versions.

The follow-up telephone survey was intended to enliven the constrained questionnaire responses with rich case study information from 10 schools. These were to cover a range of EMAs, subjects, key stages and both review types. The intention of the survey was to explore questionnaire responses in more detail, providing contextual insights that would facilitate the interpretation of the Study 4 results.
5.2 Methodology

The main methodological issues for the postal survey concerned sample selection, questionnaire development and questionnaire design. These are discussed below, after which methodological issues relating to the follow-up telephone survey are presented.

5.2.1 Questionnaire sample selection

Sample selection for Study 4 was considerably more complicated than for Study 3. Not only did seven samples have to be selected according to a range of interrelated criteria, the populations from which they were to be drawn had first to be constructed. This meant collecting data from each of the EMAs, to determine which schools had requested which types of review and to gather essential contact information. This then had to be collapsed into one national database which then had to be divided into seven separate ones. The samples were drawn from these seven populations (i.e., of all schools that had requested R2 or R3 reviews during 2000 in each of the subjects at each key stage).

It was intended to draw a sample of 250 schools from each of the six R2 populations and 500 schools from the R3 population. In fact, one of the populations (KS3 maths) had fewer than 250 schools, so (virtually) all of these were included. Similarly, KS3 English R3 had fewer than 500 schools, so most of the population was included. For the remaining populations (with more than 250 schools) a stratified sampling technique was employed to ensure that each sample was as representative as possible of all schools in its population. The samples were stratified by the following four factors:

1. **school performance** (from the 1998 primary and secondary performance tables). Schools were divided into 5 categories: from the lowest scoring 20% to the highest scoring 20% (based upon KS2 results or GCSE results).

2. **school type**. Primary schools were divided into 4 categories: Primary/Combined; Junior; Middle; and Independent. Secondary schools were divided into 7 categories: Middle deemed secondary; Secondary Modern; Comprehensive to 16; Comprehensive to 18; Grammar; Independent; and Other.

3. **region**. Schools were divided into 3 categories: North; Midlands; and South.

4. **EMA**. Schools were divided into 3 categories: AQA-S; AQA-N and Edexcel.

In fact, the sampling was further complicated by the requirement that no school should be sampled more than once, even if it had submitted review requests for more than one subject.
5.2.2 Questionnaire development

The questionnaires were designed to address a number of issues which included: methods of script checking and decision making; history of requesting reviews; impressions of reviews outcomes; motivations for requesting reviews; attitudes to the deposit; etc. Although seven versions were prepared, they all followed the same basic template and shared as much as possible in common (so as to facilitate comparison between versions). Question response formats were designed to be as closed as possible (i.e., tick-boxes), to ensure that the survey was not too time-consuming for schools to complete.

Unfortunately, the development of the questionnaires took place during the summer vacation which meant that it was not possible to pilot draft questions. However, before the summer vacation began, it was possible to contact a number of schools that had requested reviews during 1999 to establish what they considered to be the most salient issues.

A copy of each of the seven versions of the questionnaire can be found in Appendix 5.1.

5.2.3 Questionnaire distribution

Before the questionnaires were distributed, letters were sent to LEAs (which included lists of prospective schools) to seek their approval. (This is standard NFER practice to ensure that undue demands are not placed on schools which face particular problems.)

Questionnaires were distributed on the 2nd October 2000, by which time all schools should have had time to have considered their review outcomes. For both primary and secondary samples, the questionnaires were sent to schools with cover letters addressed to the Headteacher. For the secondary samples it was suggested that the Head of the relevant subject department would probably be best placed to complete the survey and Headteachers were encouraged to pass all documentation to them. For the primary sample it was suggested that the Headteacher might be best placed to complete the survey, although they were encouraged to consult with other teachers where, or if, necessary.

Reminder letters, which included a copy of the original questionnaire, were despatched on 16th October to schools that had not responded to the original survey by 13th October. The last batches of returned questionnaires were sent to the NFER Data Entry department by 27th October. Here they were double-entered, to ensure the fidelity of the data capture.
5.2.4 The follow-up telephone interviews

The follow-up telephone survey was intended to elicit qualitative information that could usefully contextualise the quantitative information generated by the questionnaires. As such, and because of the practical limitations of time and cost, there was no expectation that the results should be in any way representative of all schools. Instead, the 10 schools were selected simply to reflect a range of respondents with potentially different perspectives upon the review process. The first criterion for sample selection was that the interviewees should be distributed according to the following scheme:

1. KS2 English R2 AQA-S
2. KS2 English R2 Edexcel
3. KS2 maths R2 Edexcel
4. KS2 science R2 AQA-S
5. KS3 English R2 AQA-N
6. KS3 English R2 AQA-S
7. KS3 maths R2 AQA-N
8. KS3 science R2 Edexcel
9. KS3 English R3 AQA-N
10. KS3 English R3 Edexcel

The second criterion for sample selection was that each of the potential interviewees should have indicated her willingness to take part in the follow-up study in her questionnaire response. Many of the questionnaire respondents did so. Finally, from those that had expressed an interest, interviewees were selected to provide a range of perspectives, i.e.: small versus large schools; schools with many review requests versus schools with none; schools with reviews rejected versus schools with reviews accepted; interviewees who had lost faith in the review system versus interviewees who had not; etc..

The interviews tended to last around half an hour and were based around the interviewees’ questionnaire responses and an informal interview schedule (for guidance rather than to be followed rigidly). The schedule highlighted four main issues that would be of relevance:

1. **Background information.** For example, what were the characteristics of the schools and the respondents?
2. **2000 review requests.** For example, how were the decisions to request reviews made and what were the impressions of the outcomes?

3. **Satisfaction with the service received and the process generally.** For example, were there any aspects that especially impressed or annoyed the interviewees?

4. **Further elaboration of responses to key questions.** In particular, questions concerning preparedness to tolerate marking error and loss of faith in the external marking and review system.

The interviews were written up as individual case study reports (Appendix 5.2) and used to support and extend inferences from the questionnaires.

## 5.3 Results

### 5.3.1 Response rates

The response rates across all seven samples were good, ranging between 56% (KS2 maths) and 76% (KS3 English R3). A full response-rate breakdown is presented at the beginning of Appendix 5.3.

### 5.3.2 Reply-sample representativeness

The first task was to determine whether the samples that replied to each questionnaire were representative of the sample to which each questionnaire was sent (each of which had previously been constructed to be as representative as possible of the seven populations of schools that had requested reviews during 2000).

In order to determine whether the seven 'reply-samples' were representative of the seven 'despatch-samples', each was broken down by the four stratification factors presented earlier and the relative proportions of respondents within the categories of each factor were examined (see Appendix 5.3).

Statistical tests were used to determine whether the proportions of respondents in each category of each factor for each reply-sample matched the proportions of respondents in each category of each factor for each despatch-sample. None of the tests gave any indication of significant differences.

These results suggested that the seven reply-samples were representative of the seven despatch-samples in terms of the four stratification factors: school performance; school type; region; and EMA. It can therefore be inferred that, in terms of these four factors, the
seven reply-samples were representative of the seven populations of schools that requested reviews during 2000.

In the following discussion, the term ‘sample’ will generally be used to refer to the samples of schools that returned each version of the questionnaire, rather than the samples of schools to which each questionnaire was sent: the reply-sample rather than the despatch-sample.

5.3.3 Responses to the survey questions

In general, the questionnaires functioned effectively and as intended. The vast majority of respondents answered the vast majority of the questions that they were invited to. The following discussion will highlight the main findings that emerged. A full breakdown of responses to individual questions can be found in Appendix 5.3.

5.3.3.1 Characteristics of the respondents

The presentation of results begins by characterising the respondents. To the extent that the survey respondents’ responses can be taken as representative of all review-requesting schools, the following results help to indicate the characteristics of such schools and the key protagonists within them. Furthermore, to the extent that the Study 3 survey respondents’ responses can be taken as representative of all schools that participated in national curriculum testing during 2000, a comparison of Study 4 results with Study 3 results (for comparable questions) further illuminates the character of review-requesting schools.

5.3.3.1.1 History of requesting reviews (Q3, Q27)

During the research for Study 1, a perception amongst EMA project managers that certain schools request reviews every year was noted. Were schools that requested R2 (or R3) reviews during 2000 more likely than other schools to have requested reviews in previous years? When responses to Question 3 were explored, it appeared that they were. Across all seven samples, the percentages that had requested either R1, R2 or R3 reviews during 1999 ranged from 22% (KS2M) to 45% (KS3S and KS3ER2). This contrasted sharply with the national picture for 1999, which ranged from 6% (KS2M) to 27% (KS3E). Across samples, the percentage of the 2000 review-requesting schools that had also requested some kind of review during 1999 ranged from 1½ times, to over 4 times, the national percentage.

The results from Question 3 were also used to determine how many times, during the past three years, the review schools had requested either R2 or R3 reviews. Across samples, it
emerged that between 12% (KS2M) and 33% (KS3ER2) of schools had requested script re-marks in at least three out of four years, from 1997 to 2000.

Question 27 extended this by asking respondents to indicate the likelihood that their department would be requesting at least one R2 review in 2001. No more than one-third of respondents, in any sample, indicated that they ‘Couldn’t say’ and no more than 5% responded with either ‘Unlikely’ or ‘Definitely not’. Across samples, between one-fifth and one-half of respondents ticked the ‘Possible’ box and a similar range ticked the ‘Likely’ box. Between 3% (KS2M and KS2S) and 21% (KS3ER2) ticked the ‘Definite’ box. Over one-half of KS3 English R2 and KS3 science respondents ticked either the ‘Definite’ box or the ‘Likely’ box.

Thus, the survey elicited evidence to support project managers’ impressions that some schools repeatedly request reviews from year to year.

5.3.3.1.2 Position in school (Q29)

Across samples at key stage 2, over three-quarters of respondents were headteachers. Only around 10% of respondents referred to themselves exclusively as teachers or exclusively as assessment co-ordinators.

Across samples at key stage 3, over three-quarters of respondents were heads of subject departments. Only a very small minority of respondents referred to themselves uniquely as teachers or uniquely as examinations officers.

5.3.3.1.3 Practical marking experience (Q28)

Between 83% (KS2M and KS2S) and 88% (KS2E) of respondents at key stage 2 ticked none of the boxes of Question 28, indicating that they did not have any previous experience of external marking. At key stage 3, the comparable figures ranged from 38% (KS3ER2) to 48% (KS3S). Those few primary respondents who did have experience had generally marked for the KS2 tests. The secondary respondents with experience had typically marked for the KS3 tests or for GCSE, or for both.

As precisely the same question was asked during Study 3, a direct comparison of results was possible. While exactly the same percentage of Study 4 KS3 English R3 respondents had no marking experience as the KS3 English respondents in Study 3, 7% fewer Study 4 KS3 English R2 respondents had no marking experience. For maths and science at key stage 3, 10% and 6% fewer respondents had no marking experience.
At key stage 2, the single figure of 90% from Study 3 meant that between 2% and 7% fewer of the Study 4 respondents had no marking experience.

Overall, then, the review-requesting respondents of Study 4 had slightly more external marking experience than was apparent from the respondents of Study 3 (who represented schools in general).

5.3.3.1.4 Impressions of and attitudes towards the initial marking (Q25, Q26a)

Question 25 asked respondents to rate the general quality of the initial marking of scripts during 2000. They responded using a five point scale from 'Very poor' (scored as 1) through 'Satisfactory' (scored as 3) to 'Very good' (scored as 5). For maths and science, at both key stages, around one-fifth of respondents rated the marking as 'Poor' or 'Very poor' and more than one-third rated it as 'Good' or 'Very good'. This contrasted with ratings for English, for which 46% (KS2E), 65% (KS2ER2) and 98% (KS2ER3) were 'Poor' or 'Very poor'. The same question was asked during Study 3 where, for English, only 13% (KS2E) and 45% (KS3E) of respondents believed the marking to have been 'Poor' or 'Very poor'. For maths and science, the comparable figures were all below one-tenth for both key stages.

Figure 1. Question 26a(i). “I believe that most scripts receive the marks they deserve during the initial marking phase.”
When responses to Question 26a(i) were explored, a very similar pattern was observed between samples as was observed for Question 25, with English respondents being most negative and maths and science least. This time, the question was the extent to which respondents believed that most scripts received the marks they deserve during the initial marking phase. Respondents were asked to tick one of five boxes from ‘Strongly disagree’ (scored as 1) through ‘Undecided’ (scored as 3) to ‘Strongly agree’ (scored as 5). Across samples, the tendency was for around 5% to 10% fewer respondents to tick the ‘Strongly disagree’ or ‘Disagree’ boxes than ticked the ‘Poor’ or ‘Very poor’ boxes for Question 25.

Questions 26a(ii) and (iii) asked participants to rate the extent to which they were prepared to tolerate ‘a small amount’ of R1 and R2 marking error, respectively. In fact, responses for the two review types were very similar. Indeed, they were also very similar across samples, with around two-thirds to three-quarters of respondents ticking either the ‘Disagree’ or ‘Strongly disagree’ boxes. Only around one-fifth of respondents agreed that they were prepared to tolerate a small amount of error. Maths and science respondents, at key stage 3, were the most tolerant of all groups; but the majority of each of these still disagreed.

Study 3 asked a very similar question and received a considerably less extreme response. Here, only at KS2 did the majority of respondents (three-fifths) either ‘Disagree’ or ‘Strongly disagree’. For English, maths and science at key stage 3, only one-third of respondents disagreed with the statement that they were prepared to tolerate a small amount of marking error, while one-half agreed.

Respondents were also asked to indicate the extent to which they had lost faith in the external marking system, on the same scale from ‘Strongly disagree’ to ‘Strongly agree’. Again, the trends across samples were very similar to those for Questions 25 and 26a(i), with views being most negative for KS3 English R3 (84% agreeing), KS3 English R2 (56% agreeing) and KS2 English (25% agreeing). No more than 14% of maths or science respondents agreed. Only for KS3 English did more than 5% of respondents ‘Strongly agree’ (32% for R2 and 57% for R3).

In general, then, the review-requesting respondents of Study 4 had a less positive impression of the quality of initial marking (although no there was no evidence to determine whether this was exclusively a cause or an effect of lodging a review request).
5.3.3.2 The process of requesting a review

The second major area investigated through the questionnaire was the way in which respondents’ schools approached the review process. This was intended to determine, from a school’s perspective, both what is involved in requesting a review and how the process of requesting a review tends to be perceived.

5.3.3.2.1 How do schools check the marking of test scripts? (Q4)

When asked how they checked the initial marking of test scripts, there was a notable difference between the way in which maths or science and English respondents answered. For maths and science, at both key stages, the most frequently selected option was the ‘review of selected scripts for R2 errors’ (between 57% and 77%). For English, the most frequently selected option was the ‘comparison of expected versus obtained levels’ option (between 63% and 83%), followed by the ‘review of selected scripts for R2 errors’ option (between 60% and 68%).

Quite a substantial percentage of schools claimed to have checked all of their pupils’ scripts for R2 errors. This ranged from 15% of schools (KS3S) to 36% (KS2S and KS3ER3). This was considerably higher than when the same question was asked of a representative sample of all schools during Study 3.
The three most frequently reported response-patterns tended to be: for schools to check selected scripts for R2 errors; to check expected versus observed levels and selected scripts for R2 errors; or to check expected versus observed levels, selected scripts for R1 errors and selected scripts for R2 errors.

Finally, a small but noteworthy percentage of schools, across all subjects, claimed only to have checked expected versus observed levels. This ranged from 3% (KS3 S) to 10% (KS3 ER2).

5.3.3.2.2 How did schools perceive apparent errors? (Q8, Q9, Q10)

Questions 8 to 10 explored respondents’ impressions of the nature of the apparent errors in the marking of the scripts that they submitted for review. Question 8 began by asking whether they would describe the typical discrepancies identified as ‘small’, ‘medium’ or ‘large’. Following a pattern that has already emerged for several other questions, the highest percentage of ‘large’ errors was observed from KS3 English R3 respondents (68%), then KS3 English R2 (33%), then KS2 English (13%); in contrast, between 6% and 9% of maths and science respondents reported ‘large’ errors.

This trend was repeated when respondents were asked how many questions they typically queried the marking of in each of their review requests. The percentage of KS3 English R3 respondents that reported ‘a few’ or ‘many’ (as opposed to ‘one’ or ‘two’) was 86%; this fell to 66% for KS3 English R2. Then came KS2 English and KS3 science, with 47% and 44%, respectively; followed by KS2 science, 24%, KS2 maths, 14% and KS3 maths, 13%.

When asked to describe how the marking appeared to be in error, not surprisingly (given the findings of Study 2), virtually no respondents reported that the marking was too lenient. Across subjects, most respondents tended to record either ‘harsh’, ‘inconsistent’ or (to a slightly lesser extent) ‘inexplicable’. There were no clear patterns, though, other than there being no one option that the majority of respondents recorded (the one exception was KS2 science, for which 53% recorded ‘harsh’).

5.3.3.2.3 Who do schools involve in the process of requesting a review? (Q5, Q6)

Question 5 asked who schools involved in the decision to request a review. Generally, consistent patterns were observed across subjects within each key stage. For primary schools, these involved were predominantly headteachers (around 93% of schools), teachers (around 77% of schools), assessment co-ordinators (around 33% of schools) and senior management teams (around 24% of schools). For secondary schools, the most
frequently involved individuals were heads of departments (between 79% and 96% of schools), teachers (between 69% and 75% of schools), headteachers (between 29% and 83% of schools) and examinations officers (between 25% and 41% of schools).

The most notable difference between samples at key stage 3 was for the involvement of the headteacher. While only one-third of schools involved their heads in decisions for maths and science, this rose to three-fifths for English R2 and four-fifths for English R3.

Consistently few schools involved either governors, LEA officers or parents. Only for KS3 maths and KS3 science did a considerable percentage of schools involve pupils (20% and 13%, respectively).

Question 6 asked who was responsible for filling in the review request forms. The most frequently recorded response-pattern at key stage 2 was the headteacher alone (29% to 48% of schools) and, at key stage 3, was the head of department alone (32% to 48% of schools). Another common pattern, at both key stages, was for the head (KS2) or head of department (KS3) to complete the forms with the assistance of a teacher.

At key stage 2, across all subjects, assessment co-ordinators were involved in completing the forms in only around one-fifth of schools. At key stage 3, examinations officers were involved in completing the forms in 18% (English), 27% (science) and 41% (maths) of schools. The findings clearly indicated that the processing of review requests was typically not treated as a general administrative task, but was given special priority by heads and heads of departments.

5.3.3.2.4 How long did the process take? (Q7)

It was expected that checking pupils' scripts for marking errors might take up a considerable amount of time. Question 7 asked respondents how many person-hours were devoted to this task during 2000. Across subjects, the mean responses ranged from 3.8 hours (KS2S) to 13.2 hours (KS3ER3). However, the standard deviations associated with these means tended to be as large, if not larger. This is, perhaps, not surprising given the range of approaches to checking the marking of pupils' scripts. An alternative way of examining these data is to consider the proportion of schools that claimed to have taken longer than a specified time. If that time is taken to be 5 hours, the percentages of schools that took longer ranged from 18% (KS2M and KS3S), through 40% (KS3S), to 72% (KS3ER3).

The same question was posed regarding the time taken to prepare review requests. Across samples, the means ranged from 1.0 hour (KS3M) to 5.8 hours (KS3ER3). With a cut-off of one hour, the percentage of schools that took longer ranged from 13% (KS3M),
through 42% (KS3S), to 93% (KS3ER3). Indeed, 22% of KS3 English R2 schools claimed to have taken more than 5 hours to complete their review requests and 20% of KS3 English R3 schools claimed to have taken longer than 8 hours.

**Figure 3.** Question 7. The percentage of schools that took longer than 5 hours to check pupils’ scripts.

5.3.3.2.5 **What motivates schools to request reviews? (Q11, Q26b(x))**

As with Study 3, one important objective was to explore schools’ motivations for requesting reviews. Question 11 explored a number of possible factors. It asked respondents to rate each of five possible motivations on a five point scale from ‘1 (not at all significant)’ to ‘5 (very significant)’.

Not surprisingly, the first potential motivation – to ensure that children’s levels were as accurate as possible – was rated as significant by virtually all respondents across all samples. The mean rating was 4.9 for all samples with the exception of one 4.7.

The next possible motivation – ensuring that the school scored as highly as possible – was rated reasonably highly, but considerably less so than for the first. It received mean ratings between 3.3 (KS3ER3) and 4.1 (KS2M). While the majority of respondents, across samples, tended to tick 4 or 5, the remainder tended to be spread evenly between 3
and 2 or 1. Notably, between 14% (KS2M) and 24% (KS3ER3) of respondents ticked boxes 1 or 2, indicating that this was not a significant motivation.

Study 3 had asked a similar question (of schools in general) and the results were very similar to those that emerged from Study 4. In Study 3, the mean ratings for this question ranged between 3.3 (KS3E) and 3.9 (KS3S).

Figure 4. **Question 11b. How significant a motivation was: ensuring that the school scored as highly as possible?**

The third motivation was different from the first only in terms of emphasis. While the first concerned ensuring that children’s levels were as accurate as possible, the third concerned ensuring that children scored as highly as possible. This distinction was included because the form used during Study 3, which was most similar to the third motivation of Study 4, was annotated by a number of schools with the comment that they were not motivated to maximise pupils’ results but to ensure that results were accurate. During Study 4, ensuring that all children scored as highly as possible was given high ratings – not as high as ensuring that pupils’ levels were as accurate as possible, but higher than ensuring that the school scored as highly as possible. Across samples, the mean ratings ranged from 3.8 (KS3ER3) to 4.5 (KS2M).

An important question was whether schools had been motivated to use the review service as a way of lodging a formal complaint against the external marking system. During
Study 3, this did not emerge as a significant motivation, with mean scores ranging from 1.9 (KS3S) to 2.4 (KS3E). In fact, Study 4 responses were even lower, with a median mean rating (across samples) of 1.6. Across all samples, at least half of respondents gave this motivation an insignificant rating of 1 or 2.

**Figure 5.** Question 11e. How significant a motivation was: discrepancies between TA and test levels?

Finally, respondents were asked the extent to which observed differences between test and teacher assessment results motivated them to request their R2 review(s). Notably, there was a tendency for responses to differ by subject. For maths and science, at both key stages, the mean ratings ranged from 2.0 (KS3M) to 3.2 (KS2S). For English, the mean ratings ranged from 3.6 (KS2E) to 4.1 (KS3ER3). Recalling that, for English (but not for maths nor science), the most frequently selected checking option was the ‘comparison of expected versus obtained levels’. For English, responses to the ‘test vs TA’ motivation question were heavily skewed towards the higher (significant) rating categories; for maths and science, responses tended to be more evenly distributed across the five possible rating categories.

When the Study 3 version of this question was analysed, the same distinction between responses for maths and science versus English was not quite as apparent (although there was only one KS3 English questionnaire and the KS2 questionnaire was not specific to any subject). The mean ratings ranged between 3.3 (KS3M and KS3S) and 4.0 (KS3E).
A word of caution must be noted when comparing responses to the Study 3 and Study 4 versions of Question 11. For Study 3, respondents were asked to rate "how significant would the following be in motivating your decision" (hypothetical event). For Study 4, respondents were asked to rate "how significant were the following in motivating your school’s decision to request an R2 review this year" (actual event). A factor that might in principle have been considered a motivation, might not have actually been a motivation during 2000.

Finally, insights into what motivates schools to request reviews might come from responses to Question 26b(ix). Respondents were asked to express their agreement with the following statement: a significant number of schools use the review process more to increase their results that to see justice done for their children. Mean ratings for this question, across all samples, clustered just below the rating of 3 which corresponded to the 'Undecided' category. Indeed, for all samples, between one-third and one-half of respondents actually ticked this box. However, over one-fifth of respondents from five of the seven samples (excluding KS3ER2 and KS3ER3) agreed with the proposition. These findings were very similar to those that emerged when the same question was presented to the Study 3 respondents.

5.3.3.2.6 How smooth was the process of requesting a review? (Q14)

Respondents were presented with a series of eight statements that related to the effectiveness of procedures for requesting reviews. They were asked to respond using the five point scale from 'Strongly disagree' (scored as 1) through 'Undecided' (scored as 3) to 'Strongly agree' (scored as 5).

The first statement – it was hard to locate a copy of the Schools' Guide to External Marking – was generally disagreed with. Mean ratings ranged from 1.7 (KS2E and KS2S) to 2.1 (KS3ER2). There was a slight tendency for more agreement from secondary schools. At key stage 2, the percentage of respondents that agreed ('Strongly agree' or 'agree') ranged between 3% (science) and 9% (maths); at key stage 3, the percentages ranged between 10% (maths and English R3) to 14% (English R2).

The second statement – that the 'Review Services' section of the Schools' Guide was difficult to follow – received slightly higher mean ratings across all samples. These ranged from 1.9 (KS3S) to 2.5 (KS3ER3). Over one-quarter of KS3 English R3 respondents agreed that the Review Services section was difficult to follow. However, over two-thirds of respondents, across all seven samples, disagreed.
When asked to rate their opinion of the third statement — that there was too much paperwork to complete for the review request — the mean ratings were slightly higher again, for all samples. They now ranged from 2.2 (KS3M) to 3.7 (KS3ER2). Yet, only for KS3 English (R2 and R3) did the mean rating fall on the ‘agree’ side of the ‘undecided’ middle point of the scale. One-half of KS3 English R3 respondents agreed that there was too much paperwork, as did two-thirds of KS3 English R2 respondents. The majority of respondents in each of the other samples disagreed.

Figure 6. Question 14c. “There was too much paperwork to complete for the review request.”

The fourth statement — too much unnecessary review information was requested — was answered in almost the same way as for the third. Again, respondents for English at key stage 3 recorded the most extreme ratings, with 57% (KS2ER2) and 41% (KS3ER3) agreeing.

Responses to the fifth statement — important review information was not requested — were less in agreement, with mean ratings ranging from 2.1 (KS2M, KS2S and KS3S) to 2.8 (KS3ER3). Patterns across subjects were similar with around one-half (key stage 3) and two-thirds (key stage 2) of respondents ticking the ‘Disagree’ box.

In response to the sixth statement — the review procedures are too complicated — the majority of respondents for maths and science and KS2 English indicated ‘Disagree’. For
KS3 English, responses were almost evenly distributed around the middle ‘Undecided’ category. For both R1 and R2, one-third of respondents ticked the ‘Disagree’ category and one-third ticked ‘Agree’. Thus, views were evenly divided, but not extreme.

Figure 7.  Question 14g. “We had insufficient time to complete our review request(s).”

The seventh statement – we had insufficient time to complete our review request(s) – received mean ratings that tended to cluster just below the middle ‘Undecided’ category. They were lowest for maths and science, ranging from 2.1 (KS3M) to 2.7 (KS2M); and highest for English, ranging from 2.9 (KS2E) to 3.7 (KS3ER2). Of all the statements of Question 14, respondents demonstrated the most concern in relation to this one. For KS3 English R2 and R3, 37% and 23% of respondents, respectively, ticked the ‘Strongly agree’ box. With the exception of KS3 maths, between one-third and two-thirds of respondents across samples agreed with this statement.

Finally, a word on the number of scripts sent in for review. While questions regarding the number of scripts submitted for R1 and R2 review were mainly of use for determining the extent to which respondents’ schools had received positive or negative outcomes, the question for R3 was of interest in its own right. Question 1, for the KS3 English R3 sample, revealed different practices in the submission of R3 review requests. While the clear majority, 70% of respondents, submitted scripts for all pupils that sat the tests, one-quarter of respondents submitted only selected scripts from those that they deemed to be
poorly marked. A small minority, 5% of respondents, selected all scripts from those that they deemed to be poorly marked.

5.3.3.2.7 Thoughts on the deposit (Q14h, Q12, Q13)

The final statement of Question 14 – it was not easy to raise the deposit cheque – was prompted by comments from a number of schools during earlier studies that it was not easy to raise a deposit cheque in advance of receiving an invoice. Indeed, one of the Study 4 questionnaires was annotated as follows:

"The secretary was also unhappy about writing a cheque which might not be cashed. Apparently it is difficult for the computer system to deal with such things." [KS3 maths]

Generally speaking, though, raising the deposit had not proved to have been a problem for Study 4 respondents. At least 77% of respondents, across samples, disagreed with the statement. However, it should also be noted that, across samples, between 6% (KS2M) and 12% (KS3ER2) of respondents agreed with it. Indeed, there were respondents in all samples (except KS3M) who ticked the ‘Strongly agree’ box.

Questions 12 and 13 attempted to explore the extent to which the deposit served as a deterrent to the request of reviews. Firstly, respondents were asked whether the number of review requests that they made would have been affected if the deposit had been set at £20 per pupil (rather than the current rate of £5). Interestingly, the majority of respondents across all samples replied with ‘No’. Fewer than one-sixth of respondents, in any of the samples, replied with ‘Yes’. Having said that, a number of respondents annotated their questionnaire to the effect that they believed the value of the deposit should not be increased.

Secondly, respondents were asked whether the number of review requests submitted by their school would have been affected if deposits had been returned only for those children whose levels had actually changed. Under present arrangements, if at least one of the R2 review requests made by a school (within a subject area) results in a change of level, then the entire deposit for that school is returned. This is as true if all ten of a school’s submission are awarded higher levels, as if only one is. Clearly, then, if a school has at least one ‘dead cert.’ then submitting an additional nine ‘marginal’ scripts would mean no additional financial threat.

In fact, from annotations to a number of questionnaires it became apparent that the implication of this question had not necessarily been clear. Some respondents had commented that, even under the present system, deposits were only returned if a child’s
level had changed. Of course, this is true when only one R2 submission is made and this should have been made clearer in the question.

With the intention of excluding those respondents who may have been more likely to miss the question's implication, responses were analysed for only those respondents who had submitted more than two pupils for R2 review. Even using only these respondents, only a small minority across all samples ticked the 'Yes' box, indicating that a pro-rata return (dependent upon level change) would have affected the number of review requests made. While between two-thirds and nine-tenths of respondents were clear in replying 'No', between 13% (KS3M) and 37% (KS3S) replied with either 'Yes' or 'Maybe'.

5.3.3.3 Impression s of the service provided by EMAs

The next area addressed by the survey was the extent to which respondents were satisfied with the way in which their EMA had dealt with them.

5.3.3.3.1 Assistance given by EMAs (Q15)

Question 15 focused particularly upon those schools that had had personal contact with their EMA, either through telephone calls, faxes, emails, etc. It began by determining the prevalence of personal contact.

With the exception of KS3 English R3, the majority of respondents had had no personal contact with EMA staff, only between 12% (KS3M) and 28% (KS3ER2) of respondents had. On the other hand, one-half of KS3 English R3 respondents had had personal contact with EMA staff. Respondents who had made personal contact were asked to rate their agreement with four brief statements concerning the quality of service received. (It should be noted that smaller sample sizes, typically around 20 to 30, may affect the interpretation of the following data.)

Across all samples, the vast majority of respondents agreed that EMA staff were polite. This ranged between 85% (KS2ER2) and 96% (KS3S). No more than 5% of respondents, across any of the samples, ticked either the 'Disagree' or the 'Strongly disagree' box in response to this statement.

When asked to express their agreement or disagreement with the statement that staff were helpful, the responses were slightly less enthusiastic, but still very positive. At least three-fifths of respondents in all samples agreed that staff were helpful.
The percentage agreement fell again when asked whether the EMA staff were efficient. Yet, still, more than one-half of respondents in all samples ticked either the ‘Agree’ or ‘Strongly agree’ box.

Figure 8. Question 15b(iii). “Staff were efficient.”

Finally, similar levels of satisfaction were evident when respondents were asked to agree or disagree with the statement that staff were unnecessarily bureaucratic. More than one-half of respondents in all samples ticked either the ‘Disagree’ or ‘Strongly disagree’ box.

Generally speaking, then, respondents were happy with the quality of service provided by the EMAs.

5.3.3.3.2 The despatch of review outcomes (Q16)

The EMAs generally do not return review outcomes during the summer break, even if they have been completed. This is based on the assumption that there will not necessarily be anyone in school to receive them and the returned scripts. Question 16 asked whether schools would find it helpful if review outcome letters could be despatched, to named individuals, as and when available during the summer break. Response patterns were very similar across samples, with around one-half of respondents ticking the ‘Yes’ box, one-third ticking the ‘No’ box and one-sixth ticking the ‘Maybe’ box.
5.3.3.4 Impressions of review outcomes

One of the areas that a number of the EMA project managers were particularly keen for feedback on concerned the reception given to outcome reports. When a script is awarded a higher level, the EMAs give very little in the way of feedback. However, when a school’s request for a higher level is not granted, the EMAs prepare outcome reports which explain why. The following sections outline schools’ impressions of these reports and their views on what options should be available to them if they disagree with the EMA.

5.3.3.4.1 Reactions to outcome reports (Q19 to Q22)

Questions 19 to 22 were addressed to those schools that had had R2 (or R3) review requests turned down and, therefore, had received outcome reports explaining why. With the exception of KS3 maths and KS3 English R3, these questions were answered by at least 50 respondents in each sample (at least 23 for KS3M and at least 35 for KS3ER3).

The first question posed was whether the reasons given by EMAs were generally understood; the response categories were ‘Yes’, ‘No’ and ‘Partially’. Across most samples, around one-half of all respondents replied ‘Yes’. The one exception was KS3 English R3, where only 31% generally understood. More than two-fifths of KS3 English R3 respondents claimed not to have generally understood the reasons given; this contrasted with the other samples, for which fewer than one-quarter of respondents claimed not to have generally understood the reasons given.

When asked whether they felt that the reasons given generally addressed the concerns raised, there was somewhat more dissatisfaction. Indeed, with the exception of KS3 English R3, the percentage of respondents that said ‘No’ ranged from 24% (KS2 English) to 38% (KS2 maths). For KS3 English R3, a considerable majority, 69% claimed that the reasons given did not address the concerns that they had specified. However, with the exception of KS3 English R3, over two-fifths of respondents claimed that the reasons given either addressed, or partially addressed, the concerns specified.

Question 20 asked whether respondents generally agreed with the reasons given for turning down their review request(s). Across all samples, respondents tended not to agree. Around one-half of respondents in each sample replied ‘No’ – they did not generally agree with the reasons given for rejecting their review request(s). For KS3 English R3, this figure was 78%. No more than 26% of respondents, in any of the samples, replied with ‘Yes’.
Finally, respondents were asked to indicate whether the amount of detail included in outcome reports was 'Too much detail', 'About the right level of detail' or 'Too little detail'. For all but one of the samples, no respondents replied that there was too much detail. Between 55% (KS3 ER2) and 70% (KS2 M and KS3 M) of R2 respondents claimed that the reports contained about the right level, while 77% of KS3 English R3 respondents claimed that there was too little detail.

From these results it appeared that the feedback given to schools was not generally being received in the manner intended by the EMAs; schools’ impressions of their outcome reports were generally far less positive than might have been hoped.

5.3.3.4.2 A mechanism for appeal (Q23, Q24)

Once an EMA has reached its decision concerning the level that a reviewed script deserves, schools have no formal mechanism through which to lodge an appeal. Question 23 asked schools whether there should be a procedure through which they could lodge appeals. Overall, the response was positive, but there were a substantial number of respondents that responded 'No' – this ranged from 6% (KS3 ER3) through a median of 22% (KS2 E) to 30% (KS3 M). The percentage that responded 'Yes' ranged from 33%...
(KS3S) through a median of 52% (KS2E and KS2M) to 76% (KS3ER3). Around one-third of respondents, across samples, replied ‘Maybe’.

When asked whether, if there had been a formal procedure during 2000, they would have considered using it, responses again spread across the three categories of ‘Yes’, ‘No’ and ‘Maybe’. Affirmative responses were highest for English: of the 77% of KS3 English R3 respondents that answered the question, over one-half said that they would have; of the 88% and 84% of KS3 English R2 and KS2 English respondents, respectively, two-fifths said that they would.

5.3.3.5 Impressions of the review system

The final area covered in the survey related to general impressions of the review system. These were answered through Question 26b, which presented a list of ten statements and asked respondents to indicate the extent to which they agreed or disagreed with them. Once more, this employed the five point scale from ‘Strongly disagree’ (scored as 1) though ‘Undecided’ (scored as 3) to ‘Strongly agree’ (scored as 5).

5.3.3.5.1 How fair is the review system? (Q26b(i), Q26b(ii), Q26b(iv))

Figure 10. Question 26b(i). “The review system is generally fair.”

Three statements within Question 26b addressed issues of fairness, none more directly than the first statement: the review system is generally fair. With the exception of KS3
English, the mean ratings across samples were extremely similar, either 3.7 or 3.8. The mean rating for KS3 English R2 was 3.4 and, for KS3 English R3, was 3.1. Thus, across all samples, the mean ratings were towards the agree side of 'Undecided'. In fact, the majority of respondents for all samples, except KS3 English R3, were agreed that the review system is generally fair. More than one-fifth of KS3 English respondents indicated that they disagreed with this proposition.

Very similar patterns of responses across subjects were observed in relation to the second statement: the review system is sufficiently independent. Across all samples, the mean ratings were slightly down, but they were still at least 3.0. With the exception of KS3 English, the majority of respondents for all samples agreed that the review system was sufficiently independent. Only one-third of KS3 English respondents (both R2 and R3) were in agreement, while around one-fifth disagreed.

Finally, Question 26b(iv) stated: the External Marking Agencies are sufficiently accountable. Once more, across all samples, the mean ratings were slightly down and, for KS3 English, they now fell below 3.0. Indeed, only for KS3 maths were the majority of respondents in agreement with the statement. For KS3 English, the largest percentage of respondents ticked either the 'Disagree' or the 'Strongly disagree' box (46% for R2 and 60% for R3). For the remaining five samples, more respondents agreed that the EMAs were sufficiently accountable than disagreed; however, between 34% (KS2S) and 41% (KS2E) were undecided.

Study 3 had comparable questions relating to the 'fair' and 'independent' statements. Moreover, the results that emerged from Study 3 were very similar to those that emerged from Study 4. If any difference was to be highlighted, then it would be a trend for fewer responses in the 'Undecided' category in relation to the statement: the review system is generally fair. The mean ratings, however, were still comparable.

The results from Study 4 therefore indicated that the review-requesting schools were generally satisfied that the review system was fair, sufficiently independent and sufficiently accountable; however, contrary views were expressed by sizeable percentages of each of the samples, particularly for English at key stage 3.

5.3.3.5.2 How transparent is the review system? (Q26b(iii))

It is frequently said that, for an assessment system to be trusted, it must be transparent. While it is not always clear exactly what is meant by this, it generally carries the connotation that those with an interest in understanding the system are able to. Question 26b(iii) presented the statement: the review system is sufficiently transparent. As for the
previous four questions, respondents in all samples answered similarly, although the KS3 English respondents were generally more negative. Around one-third of KS3 English respondents agreed with the statement, compared with around one-half of respondents from the remaining 5 samples. Around one-tenth of respondents disagreed with the statement, for all samples except the KS3 English ones, where around one-third disagreed.

5.3.3.5.3 How effective is the review system? (Q26b(v), Q26b(vi))

Two questions presented statements that were intended to explore the confidence that Study 4 participants placed in the review system (having just been through it).

Figure 11. Question 26b(vi). “I have lost faith in the review system.”

The first statement – the review system generally ensures that children receive the correct national curriculum level for their work – was reacted to in a generally positive manner, by all respondents except those from the KS3 English samples. For the maths and science samples, the percentages that agreed with the statement ranged from 68% (KS2S) to 81% (KS3M); and, from the KS2 English sample, 59% agreed. However, no more than 43% of respondents from the KS3 English samples agreed and no fewer than 30% disagreed. For the remaining five samples, no more than 14% disagreed.

Putting the question in a more extreme format, respondents were asked to indicate the extent to which they agreed with the statement: I have lost faith in the review system.
With the exception of KS3 English R3, the majority of respondents in all samples disagreed. Indeed, at least three-quarters of maths and science respondents disagreed. More worryingly, at least a small minority of respondents from all samples agreed. While this figure was no higher than 8% for maths, science and KS2 English, it reached 21% for KS3 English R2 and 35% for KS3 English R3.

5.3.3.5.4 How might fewer reviews be facilitated? (Q26b(vii) to (ix))

The remaining elements of 26b explored the extent to which respondents believed that fewer reviews could be facilitated through particular interventions. These concerned the mark schemes, marker training and marker vigilance.

Perhaps not surprisingly, the statement that was agreed with the most was: there would be fewer review requests if markers were more thorough. The majority of respondents in all subjects agreed with this statement. Indeed, for KS3 English, the majority of respondents ticked the ‘Strongly agree’ box. By way of contrast, fewer than one-quarter of respondents from the remaining five samples (including KS2 English) ticked the ‘Strongly agree’ box.

Many respondents also agreed with the statement: there would be fewer review requests if markers were better trained. Indeed, for KS3 English, there was even more agreement than for the last statement: 55% (R2) and 65% (R3) ticked the ‘Strongly agree’ box, while 80% (R2) and 91% (R3) ticked either the ‘Strongly agree’ or the ‘Agree’ box. For the KS2 English and the KS3 science sample, the picture was somewhat less extreme, with 61% and 58% in agreement, respectively. Only for the KS3 maths sample did the largest number of respondents (48%) tick either the ‘Disagree’ or the ‘Strongly disagree’ box.

The final statement – there would be fewer review requests if mark schemes for current tests were better constructed – was not responded to as extremley by the KS3 English respondents. Indeed, response patterns were similar across all English and science samples. Still, though, the majority of respondents in each of these samples agreed with the statement, indicating that they believed that mark scheme development could be improved. Once again, only for KS3 maths was the largest number of respondents (43%) in disagreement.

5.3.3.6 Differences between EMAs

Results for a number of questions were broken down further to explore the extent to which the EMA responsible for marking and reviewing respondents’ scripts may have affected survey responses. When interpreting the following results, it is important to bear in mind that some of the analyses were based upon small sample sizes. This would make
it harder to detect significant differences between EMAs even if they genuinely existed; indeed, small sample sizes occasionally meant that statistical comparison was not deemed valid at all. Moreover, as discussed during Study 3, any comparisons between EMAs should be treated with caution: schools are not allocated to EMAs on a random basis, but on a regional basis; thus apparent differences due to the impact of an EMA might feasibly be re-interpreted as due to general characteristics of respondents within different regions.

5.3.3.6.1 Assistance given by EMAs (Q15)

Question 15a asked respondents whether they had had any personal contact with their EMA. When responses were compared between EMAs, significant differences were observed only for the KS3 English R3 sample. This suggested that a higher percentage of the Edexcel respondents had had personal contact with EMA staff than either AQA-S or AQA-N respondents.

Question 15b explored impressions of the service given by staff in each EMA, but only for those respondents that claimed to have made contact. This, of course, reduced the sample sizes considerably, meaning that there were typically fewer than 15 respondents representing each EMA within each of the seven samples (although around twice as many for KS3ER3). As such, any lack of evidence of trends between EMAs should be treated with some caution. The comparisons for Question 15b were based on mean ratings from a re-scoring of the full five point scale to a four point scale that did not include the ‘Undecided’ category.

Across the four statements of Question 15b, only the third — staff were efficient — led to significant differences between EMAs. Moreover, significant differences were only observed for the KS2 English and the KS3 English R3 samples. In both of these cases, AQA-N received a significantly higher mean rating than Edexcel. The intermediate mean rating for AQA-S was significantly different from neither. This suggested that AQA-N respondents in the KS2 English and KS3 English R3 samples had a better impression of the efficiency of their EMA than Edexcel respondents.

5.3.3.6.2 Reactions to outcome reports (Q19 to Q22)

The four questions that concerned impressions of outcome reports were also analysed separately by EMA. Across all seven samples, for each of the four questions, only one comparison emerged as significant. From Question 21, it appeared that AQA-S respondents from the KS3 science sample were more likely not to feel that the reasons given in their outcome reports generally addressed the concerns that they raised.
5.3.3.6.3 Impressions of the external marking system (Q25, Q26a(iv))

Various comparisons of mean ratings for Question 25 revealed no significant differences between EMAs in respondents' ratings of the quality of initial marking. However, significant differences between the seven samples were clearly identified. These differences have been described earlier.

When responses to the statement 'I have lost faith in the external marking system' were analysed by EMA, significant differences were observed for the KS2 English and KS3 science samples only. For KS2 English, the mean rating for AQA-N was significantly lower than the mean ratings for both AQA-S and Edexcel (which were not significantly different). This suggested that fewer AQA-N respondents had lost faith in the external marking system for KS2 English than AQA-S or Edexcel respondents. For KS3 science, the mean rating for AQA-N was significantly lower than the mean rating for AQA-S; the intermediate mean rating for Edexcel was significantly different from neither. This suggested that fewer AQA-N respondents had lost faith in the external marking system for KS3 science than AQA-S respondents.

5.3.3.6.4 Impressions of the review system (Q26b(v), Q26b(vi))

Finally, the two questions that addressed general impressions of the effectiveness of the review system were explored by EMA. The first related to the statement 'The review system generally ensures that children receive the correct National Curriculum level for their work'; the second related to the statement 'I have lost faith in the review system'. No significant differences were observed between responses to the first of these statements. Significant differences between responses to the second were only observed for the KS3 English R3 sample. Here the mean rating for AQA-N was significantly lower than the mean rating for both AQA-S and Edexcel (which were not significantly different). This suggested that fewer AQA-N respondents had lost faith in the review system for KS3 English R3 than AQA-S and Edexcel respondents.

5.3.3.7 Other issues

Question 31 asked respondents if they would like to raise any additional issues. As for the same question in Study 3, the vast majority of respondents took the opportunity to highlight particular concerns with the present arrangements. The following sections reproduce some of the more salient or more interesting comments. These did not always come in response to Question 31; others came in response to Q27b, or as annotations to specific questions, or in separate cover letters. The following comments are not intended to be representative in any way.
5.3.3.7.1  The quality of initial marking

"Markers are expected to apply a mark scheme to answers that are not always black and white – for which they receive dubious [...] training and pay that at best works out at £7 an hour. Don’t be surprised if corners are cut!!" [KS2 English]

"All 5 marks I queried were changed. It seemed that the marker had not sufficient understanding of science or had marked far too quickly. There were very many alterations on the mark sheet which implied lack of attention to detail. Highly likely that this will happen again.” [KS2 science]

"We are profoundly dissatisfied with the operation of these tests. I am consequently very keen to take part in any survey, telephone or otherwise, which can highlight the scandal (it is becoming nothing less) of incompetent markers seriously underrating and misjudging pupils’ work. This year (which is the fourth consecutive year we have disagreed with the test levels) 100 out of 109 boys had their marks changed, some by over 40%.” [KS3 English]

5.3.3.7.2  The script checking process

"Perhaps you should know that: when scripts arrive back they are sorted into maths sets. Teachers then return them to the pupils who are encouraged to total their own marks. Staff then go through the papers with the pupils i.e. the pupils check their papers. Pupils who are near (but just below) the next level are encouraged to check they have not been under marked. Any other marking errors we ignore because the level won’t be changed. If a borderline pupil finds (or thinks they have found) the necessary mark(s) to change the level the teacher checks the mark scheme.” [KS3 maths]

5.3.3.7.3  The importance of accuracy

"Inaccurate marking is unacceptable both for the child and the school. If schools are to be judged on these results they must be accurate. Our local secondary school uses the marks to set the groups, the wrong mark could have a detrimental effect on the child’s Y7 education. 3% is a wide margin of error, the script was upgraded by 3%.” [KS2 maths]

5.3.3.7.4  The decision to appeal

"One year we had SAT results which were massively inflated. Of course we did not appeal. This year, the results were massively depressed and were changed by
large amounts following appeal. We cannot trust the process at all.” [KS3 English]

5.3.3.7.5 The amount of work required of teachers

“Where several errors are discovered on papers it would be helpful to send all scripts back for checking. My staff have other work to do and are not paid for marking. They might consider this if they were paid.” [KS2 maths]

“The main problem in school is how long it takes to prepare, at a very busy time of year. Whilst you have asked how many hours it took, and this looks reasonable, it was all my free time for 4 days and I only needed 16 to be looked at.” [KS3 English]

5.3.3.7.6 Changing attitudes to the system

“This year we had over 10% writing and reading papers that we strongly felt deserved a higher level so we requested a review. Previously, the year 6 teachers have accepted that children sometimes perform differently in test conditions and it is impossible to check 90+ papers thoroughly. In future, we shall ensure that time is given to check more papers, regardless of expected level, as this year’s experience has left us doubting levels awarded to children.” [KS2 English]

“This is the first year we checked papers so thoroughly. Should we have done it before instead of having complete blind faith in the system? We usually check the mark sheets (yellow) but now check everything.” [KS2 science]

5.3.3.7.7 The need for a formal complaint mechanism

“In our sample there were several mistakes in the marking, however these did not affect the level apart from two. I feel there should be a system whereby comments could be made if a marker is making a lot of mistakes which are not justifiable due to professional judgement, but are just due to poor marking.” [KS3 science]

5.3.4 The ten case study telephone interviews

The telephone interviews were intended to provide more qualitative detail concerning schools’ practices in relation to, and views of, the review system. They are presented in Appendix 5.3 and should be read in full to achieve a thorough understanding of the issues that arose. The interviews were not intended to represent the views and practices of all schools and the sample was not selected with this in mind. Instead, the intention was to support a clearer understanding of the range of different perspectives and to comprehend
the subtlety of the issues at hand in more depth. The following sections highlight a number of key issues that arose during the discussions.

5.3.4.1 The importance of correct marks

One of the clearest messages that came through was the level of importance that the interviewees attached to accurate test levels. They highlighted a variety of reasons why it was deemed important to ensure accurate results, including factors related to schools, teachers, pupils and parents. From the perspective of a school, the importance of performance table standing was emphasised (CS10) as well as the use of test results for value added and target setting purposes (CS7). In addition, from the primary perspective, it was considered important that correct test levels were passed on to secondary schools to ensure that pupils were set appropriately and that their teachers held appropriate expectations for them (CS4).

From the perspective of pupils, accurate test levels were deemed to be important to nurture their self-esteem (CS4) and to facilitate appropriately high expectations (CS9). Moreover, it became clear how upset pupils and parents can become in the face of inaccurate test levels (CS8) and how hard it can be to undo the damage of an initially inaccurate report (CS6). Finally, from the perspective of teachers, it was noted that erroneously low distributions of scores can ‘outrage and distress’ (CS5). Indeed, the potential impact on performance-related pay was also discussed (CS10).

Interestingly, though, the opposite view was also expressed – at least with respect to the key stage 3 tests – that the results do not matter enough to “waste time on” (CS5).

5.3.4.2 The degree to which error is tolerated

Reasons such as those expressed above help to explain the positions that the interviewees adopted in relation to the tolerance of clerical errors and mark scheme application errors. For example, one interviewee felt that the political and personal consequences of error were so high that there should be no leeway for either clerical or mark scheme application error (CS3).

Some interviewees’ justifications of tolerance related not to the consequences of error, per se, but to the transparency of error. For example, it was proposed that clerical error could be tolerated to some extent because it is easy to spot and correct, while mark scheme application needs to be absolutely reliable because it is not (CS2 and CS8). Other interviewees appeared to view the issue from the alternative perspective of what can reasonably be expected of markers; they were not prepared to tolerate any clerical error but were prepared to accept a small amount of mark scheme application error (CS7).
Others suggested that they were prepared to accept a small amount of error, but only when pupils’ levels were not affected (CS8).

5.3.4.3 Checking pupils’ scripts

The case studies also helped to reveal the considerable range of practices for checking the accuracy of pupils’ levels. Initial steps in this process ranged from comparing expected versus observed results (CS1, CS3, CS9, CS10), through the scrutiny of promising ‘just below borderline’ scripts (CS7), through the scrutiny of all scripts (CS6), to getting pupils to check their own results (CS8). Even the checking of observed versus expected results took on different forms, for example, comparing mark sheet results against ‘gut knowledge’ of pupils’ abilities (CS1) or against teacher assessment results (CS3, CS9) or against expected distributions of results (CS10).

In a number of cases, numerous staff were involved in checking pupils’ scripts (CS3, CS5, CS9, CS10) and there were often a number of stages to the checking process.

5.3.4.4 The amount of work required

Interviewees often highlighted the amount of time and effort that was required to check scripts and to complete review request forms. There was some evidence that the amount of time and effort required supported a decision not to follow submission procedures to the letter (CS9). It was widely felt that the amount of time and effort required was a major disincentive to submitting review requests (CS2, CS5, CS7, CS9) – that it would mean ‘more work for staff already overworked’ (CS4). One KS3 English interviewee’s school had begun on an R3 request, but had ‘lost interest’ after the process had become too time-consuming (CS5).

It was noted that the significant amount of time and effort involved, often from teams of staff rather than individuals, can generate strong feelings of disappointment when review outcomes are not as hoped (CS3).

5.3.4.5 The deposit

There was some feeling that the review system generally was set up so as to put schools off from requesting reviews, in order to keep apparent error rates low; the deposit was considered to be one key component of this and raising it was considered to be ‘unethical’ (CS2, CS9) and ‘would only serve to increase teachers’ sense of alienation from the process’ (CS2).
The was evidence of widely differing views on the impact of the deposit. In some schools it served as a major deterrent (CS2, CS4) while in others it was of no significance (CS1, CS5).

5.3.4.6 Understanding the system

A number of interviewees raised the issue of depth of understanding of the review system and explained that a level of ‘inside information’ can considerably facilitate the process (CS7, CS9, CS10). One felt that the secret to the effective checking of scripts was knowing where markers were likely to go wrong (CS7) and believed that most appeals occurred when someone inside the school has inside knowledge of the system.

5.3.4.7 Independence and transparency

A lack of faith in the independence of the system was referred to by a number of interviewees (CS3, CS4, CS5, CS7, CS8). One explained this as follows: ‘They [the EMAs] stick up for their markers in the same way I stick up for my staff’ (CS3).

In addition, concern was voiced that the system and decisions arising from it are not transparent (CS2, CS5, CS7, CS8). One interviewee added that this weights the system against those without marking experience (CS7).

5.3.4.8 Customer satisfaction

There was evidence that this lack of transparency extended to the reasons given for turning review requests down (CS3, CS7, CS8, CS10). It was felt, for example, that reasons given were ‘nit-picky’ (CS3) or that they addressed marginal points (CS7). The need for a formal appeal mechanism was not simply raised (CS1, CS5, CS9), some interviewees explained that their school would have used it this year (CS3, CS7, CS10).

Finally, as distinct from the review process, a number of schools suggested that there ought to be an additional avenue through which to raise concerns over marking. It was felt that there should be a formal mechanism through which concerns with marking could be raised that did not necessarily relate to potential level changes (CS1, CS5, CS8).

5.4 Discussion

The aims of Study 4 were to explore how schools used the review system during 2000 and to determine schools’ views on its effectiveness. These aims were achieved successfully and the following discussion will consider the most important findings that emerged.
5.4.1 Characteristics of schools that requested reviews

One of the central evaluation questions was whether the review system operated fairly during 2000; and an important component of that question was whether it was being used appropriately. If, for example, schools that should have requested reviews did not – or if schools that should not have requested reviews did – then this would suggest that the system was not operating as fairly or effectively as it might.

Study 4 began by considering whether those schools that had requested reviews during 2000 had a similar profile to the population of all schools that participated in the national curriculum tests during 2000. If marking errors were distributed in a predominantly random fashion across markers then we might expect this to be at least approximately true.

5.4.4.1 Did review-requesting schools protest too much?

The first question was whether there was a tendency for some schools to request reviews year-in-year-out. Indeed, the percentages of the review-requesting samples that requested reviews during 1999 were all higher than the corresponding national percentages. If marking errors are distributed somewhat randomly from one year to the next, then this might suggest one of three conclusions: either that some schools are requesting reviews more frequently than they should; or that some schools are not requesting reviews enough; or both.

Alternatively, reviews might not be distributed in a predominantly random fashion from one year to the next; this might occur, for example, if markers typically remain of the same standard from one year to the next and are typically allocated to the same schools. In this situation, some schools would inevitably receive poorer than average marking year-in-year-out, which would inevitably lead to a higher probability of receiving review requests from them. As such, allocating markers to the same schools from one year to the next could pose a potential threat to the professional perception of the quality of external marking – at least for a sizeable sample of schools. Having said that, computer programs are used by the EMAs to support a relatively random allocation of markers to schools; this means that schools would generally not receive the same marker from one year to the next.

5.4.4.2 Did review-requesting schools understand the system sufficiently?

Study 2 revealed that a very substantial percentage of review requests were rejected during 2000. Indeed, this was even more true for key stage 2 than for key stage 3. Was it
possible that the review-requesting schools frequently failed to understand the review process sufficiently? One index of understanding might be the extent to which respondents had external marking experience. This would not be an absolute test, however, as other influential members of staff might have had experience; yet it might still provide a possible clue.

The first significant finding was that far fewer respondents from the key stage 2 samples had any marking experience than respondents from the key stage 3 samples. As, from Study 2, there were many more successful review outcomes at key stage 3, it may be that marking experience does have some impact upon the frequency of ‘valid’ review requests (although see the discussion of Study 2 which cites alternative explanations of this effect).

Interestingly, though, the review-requesting samples from Study 4 tended to have slightly more external marking experience than the samples of all participating schools from Study 3 (and the Study 4 case studies provided further evidence that those with ‘inside knowledge’ of the system might be more likely to request reviews). Thus, it would be hard to argue that the review-requesting samples, on average, had less understanding of review processes than the general population of all participating schools.

5.4.4.3 Did review-requesting schools have a lower tolerance for marking error?

When scripts are marked by hand, rather than by machine, marking error is unavoidable. Steps can be taken to reduce the likelihood of error and checking procedures can be implemented to ensure that errors are detected before scripts are returned to schools. Yet it is unlikely that all marking errors will be eliminated.

If schools differed in the extent to which they were prepared to tolerate a small amount of marking error then this might affect the likelihood of their submitting review requests. Although no obvious differences were detected between review-requesting respondents from the seven Study 4 samples, differences were observed when they were compared with the Study 3 respondents.

The Study 3 respondents, representing all participating schools, were considerably less extreme when asked the extent to which they were prepared to tolerate a small amount of marking error. The majority of key stage 3 respondents from Study 3 (across English, maths and science) were prepared to tolerate a small amount of marking error, although the majority of key stage 2 respondents were not. In contrast, the majority of Study 4 respondents, across all seven samples, were not prepared to tolerate a small amount of marking error.
This suggests that, particularly at key stage 3, there may have been a substantial number of schools that do not request reviews - even though their scripts contained errors.

5.4.4.4 Did review-requesting schools have a low impression of the initial marking?

Was it simply that review-requesting schools were less tolerant of marking error, or was it also that their pupils' scripts were perceived to have contained more errors? This was an issue that could only be addressed indirectly, through an analysis of respondents' impressions of the quality of initial marking. Across subjects and key stages, the Study 4 respondents' impressions tended to be somewhat more negative than the Study 3 respondents' impressions. This gave some indication that the submission of review requests was associated with a worse perception of the quality of initial marking (rather than simply reflecting a lower threshold for tolerance of marking error).

(At this point, it is worth emphasising that the foregoing conclusions have, inevitably, been somewhat speculative and have tended to ignore problems such as reverse causation, i.e., that the process of requesting a review itself may have had an impact on the perceived quality of initial marking. Possibilities such as these must always be borne in mind.)

5.4.4.5 Were review-requesting schools particularly motivated by certain factors?

While 'ensuring that the school scored as highly as possible' was rated as a significant motivation in making their review request(s), there was no evidence to suggest that review-requesting schools were more motivated by this factor than schools in general.

Likewise, Study 4 respondents were not more motivated than Study 3 respondents by the desire to lodge a formal complaint against the system; indeed, if anything, they were less motivated to.

5.4.2 The process of requesting reviews

Study 4 sought to evaluate the effectiveness of procedures for requesting reviews and to explore in more depth how schools undertook them.

5.4.2.1 How did schools check scripts?

Study 4 supplemented the results of Study 3 by indicating how schools that actually requested reviews during 2000 checked the marking of their pupils' scripts. It is perhaps not surprising that there were considerably higher numbers of review-requesting schools that checked their pupils' scripts very thoroughly.
To the extent that schools are more or less diligent in checking for marking errors, individual pupils will be advantaged or disadvantaged. Moreover, to the extent that schools tend primarily to submit reviews to increase pupils’ levels (see Study 2), the less diligent the script checking process, the more disadvantaged pupils (and schools) will be.

It was interesting to note the higher prevalence of checking observed versus expected results for English. It is not clear exactly why this should have been so. However, it is important to note that when expectations relate directly to teacher assessment, then the reliability of teacher assessment results will impact upon perceptions of the quality of initial marking; to the extent that this was more prevalent for English, this would have had more of an impact.

A further point is worthy of mention: the fact that between 3% and 10% of schools claimed only to have checked observed levels against expected levels. Recalling that the procedure for requesting R2 reviews involves specifying exactly where errors lie, this seems somewhat strange. Either there must have been a considerable number of schools that requested re-marks stating simply that the children whose scripts were submitted could not have achieved such a low level, or responses to this question were not entirely accurate.

When asked to indicate the amount of time it had taken to check scripts the dispersion of answers reflected the range of approaches adopted. While the lowest mean response, across samples – nearly four hours – may not sound like a great deal, the case studies and questionnaire annotations helped to contextualise this, for example, “all my free time for 4 days”. Many respondents found the amount of time required to check scripts and prepare review cases prohibitive and there was evidence of resentment. The discontent was highest for subjects where the mean number of review requests submitted was highest (i.e., English); that is, the most pressure was experienced by those apparently worst affected by marking error.

This is an important point to note. If the review system is to be fair, then it must be equally accessible to all. If the task of preparing a review submission was deemed to be unreasonably demanding for precisely those schools worst affected by marking error, then the system would clearly be failing in a crucial respect.

5.4.2.2 How do schools request reviews?

Despite the fact that primary schools appoint assessment co-ordinators and secondary schools appoint examinations officers, these individuals are only infrequently involved in either deciding to request reviews or in completing the review request forms. If critical
information, particularly the *Schools’ Guide*, tends to reside with examinations officers and assessment co-ordinators, rather than with the heads of department or headteachers, then problems might arise that could otherwise be avoided.

In fact, locating a copy of the *Schools’ Guide* had proven to be a problem for a small minority of the Study 4 schools – particularly at key stage 3 where more than 10% of schools were affected. Clearly, then, there is work to be done to ensure that critical information is targeted at the most appropriate recipient. Reassuringly, once located, the *Schools’ Guide* was generally not considered to be difficult to follow. The area most in need of attention appeared to be the R3 review section.

Similarly, responses to questions concerning the amount of paperwork, the complexity of review procedures and the amount of information required, were reasonably reassuring. For most of the seven samples, only the minority of respondents expressed particular concern. However, as explained earlier, this should not necessarily be taken as indicative of the effectiveness of procedures. For example, if those who were faced with the most serious marking errors were those for whom requesting a review involved too much paperwork then this would constitute a potential threat to the fairness and effectiveness of the system. In fact, the amount of paperwork was most frequently seen as ‘too much’ by respondents in the English samples – those samples in which schools submitted the highest number of review requests on average.

5.4.2.3 The R3 submission

A quick mention should be given of the different strategies that schools use when submitting the work of a ‘group’ of pupils for R3 review. When discussing intended practice with EMA project managers, it was hard to determine whether there was a clear policy on whether all scripts from a school’s cohort should be submitted, or only a sample. Question 1, for the KS3 English R3 sample, established that seven-tenths of schools submitted the work of all pupils. The remainder selected only a sample – the minority of these selected all of those deemed to be poorly marked while the majority selected only some of the most poorly marked.

The implication is clear: if a school is permitted to submit only some of the pupils from its cohort then it is at liberty to submit only those with perceived marking errors that fall close to the upper limit of a level boundary. That is, it is free to submit all of those that, through bad marking, have led to inappropriately low levels, but none of those that have led to inappropriately high levels. It may be that up to one-quarter of schools adopt this tactic. This is clearly unjust and is likely to result in inappropriately high school
performance profiles for some schools – the more inconsistent the initial marking, the more inappropriately high the post-review performance profile is likely to be.

5.4.2.4 The deposit

While raising the deposit cheque was generally not felt to be a problem, some schools reported that it was. In the light of this it may be worth at least considering an invoice-based system, such as used at GCSE/A/AS (see Study 1).

When asked whether a four-fold increase in the cost of the deposit would affect the likelihood of their school submitting review requests, a surprisingly large number of respondents said not. Whether this would really translate into practice is another question. Note that the Study 3 version of this question (which had different response categories) revealed that raising the deposit to £20 per pupil was perceived as a significant discouragement to the request of a review.

The case studies helped to contextualise Study 4 responses to the deposit rate question. It became clear that, to some schools, money really did appear to be no object as far as review requests were concerned. For others, though, the deposit was clearly a strong disincentive – even at its present rate. To the extent that increasing the deposit might impact upon certain schools far more than others, a question mark is raised concerning the fairness of doing so.

5.4.3 The service provided by the EMAs

Generally speaking, of those schools that had made contact with EMA staff during the review process, most were happy with the level of service that they had received; the majority agreed that staff were polite, helpful and efficient and disagreed that they were unnecessarily bureaucratic. The only differences between EMAs were in response to the ‘efficiency’ statement for KS2 English and KS3 English R3, where AQA-N respondents were more impressed than Edexcel respondents.

The EMAs generally follow the practice of not despatching review outcomes (i.e., ones that are completed during the summer break) to schools before the beginning of the autumn term. When schools were consulted on this, the majority indicated that they would prefer to have results despatched – to named individuals – as and when available. It would, therefore, be worth exploring the feasibility of this option, at least for schools that would appreciate the service.
5.4.4 Reactions to review outcomes

As indicated earlier, a number of EMA project managers expressed particular interest in knowing how well received their review outcome reports were. The preparation of these reports can consume a considerable time and it would be reassuring for them to know that they were useful.

5.4.4.1 Reactions to the ‘reasons for rejection’ presented in outcome reports

Unfortunately, survey responses were less than encouraging. Even for the most ‘forgiving’ of the seven samples – KS3 maths – only just over one-half of respondents replied that they understood the reasons given by their EMA for rejecting their review request(s). For the KS3 English R3 sample, only one-third claimed to have understood. When asked whether they felt that the explanations given by their EMA had addressed the concerns that they had specified, the responses were even less positive.

The least positive response came when respondents were asked whether they agreed with the reasons given. Not even one-third of any sample agreed with this proposition and at least two-fifths of respondents, across samples, explicitly rejected it.

While it would be easy simply to dismiss these survey responses as little more than ‘sour grapes’ it would be unwise to do so. At the very least they indicate that the outcome reporting process is not having its intended effect of educating schools. The purpose and nature of the outcome reporting process is therefore very much in need of reconsideration. Indeed, if schools really do tend to see outcome reports as little more than face-saving ‘nit-picking’ on behalf of the EMAs, then it may well be that they are doing far more harm than good.

There is a serious issue here. If there was an element of ‘sour grapes’ in the reactions that schools had to the rejection explanations – moreover, if this is an understandable, perhaps even inevitable, human characteristic in such circumstances – then it might be better not to prepare such reports at all. Whether there are alternative ways of achieving the intended educative goal is an open question.

Finally, when responses were explored by EMA, the only significant difference to emerge was for KS3 science where fewer AQA-S respondents felt that the reasons given addressed the concerns raised. Whether this was related to the use of a standard response format was not clear.
5.4.4.2 A mechanism for appeal

Although respondents generally agreed that there ought to be an appeal process, this was by no means unanimous. Indeed, between 6% and 30% stated that there should not be. Once again, though, it is important to explore the 'worst case' scenario which, in this case, was KS3 English R3, where over three-quarters of respondents indicated that there should be an appeal process. Technically speaking (from the perspective of marking reliability), a formal appeal process might be more of a necessity for English than for maths.

Yet the possibility of an inappropriately high appeal submission rate would have to be considered, particularly bearing in mind the number of respondents that did not agree with the reasons why their review requests were turned down. Indeed, two-fifths or more of respondents in each of the three English samples indicated that they would have considered lodging an appeal during 2000.

5.4.5 Impressions of the review system

A number of questions from the survey got straight to heart of the 'customer satisfaction' issue, by asking directly for impressions of the review system.

5.4.5.1 Fairness, independence, transparency and accountability

Across all of the seven samples, only for KS3 English was there a substantial body of opposition to the statement that the review system is generally fair. This was true for more than one-fifth of KS3 English respondents (R2 and R3), but only for around one-twentieth of respondents in the other samples. However, even for KS3 English R3, just under one-half of respondents agreed that the system was fair.

Similar, although generally somewhat more negative, findings emerged with respect to impressions of independence, transparency and accountability. Even for impressions of accountability, though, (which were least positive) the percentages of respondents that agreed with the statement were higher than the percentages that disagreed, for all samples except the two KS3 English ones.

While there was general agreement concerning the perceived fairness, independence, transparency and accountability of the review system, the dissenting voices should not be forgotten. As noted in the discussion of Study 3, KS3 English presents the external marking system with particular problems; these clearly extend to review services and are of such a proportion that they deserve somehow to be addressed.
5.4.5.2 Confidence in the system

Similar messages concerning the particular problems for KS3 English emerged when respondents were asked to indicate their confidence in the review system. Only a minority of the KS3 English R2 and R3 respondents agreed with the statement ‘the review system generally ensures that children receive the correct national curriculum level for their work’ (while the majority of respondents in the five other samples agreed). Likewise, only for the KS3 English R3 sample did the majority of respondents agree with the statement ‘I have lost faith in the review system’. This contrasted with fewer than 8% of respondents from the maths and science samples and from the KS2 English sample.

It should be stressed that these impressions did not simply correspond to differences in the proportions of successful review requests between samples. For example, from Table 44 of Study 2, 70% of KS3 English R2 requests resulted in level changes. This compared favourably with the 76% for KS3 maths and 62% for KS3 science; it compared even more favourably with percentages for key stage 2, which ranged from 43% to 55%.

Exactly why the KS3 English samples should have had markedly less confidence in the review system than the other samples is not entirely clear. However, it seems likely to be at least somehow related to a more general discontent with the quality of marking for KS3 English. The fact that there are higher numbers of R2/3 review requests for KS3 English, in combination with a relatively high success rate, does support this general concern with the quality of initial marking (even if concern with the quality of re-marking is not similarly supported). Responses to other questions also indicated that the KS3 English samples, in particular, felt strongly that fewer review requests would result if markers were better trained and if they were more thorough.

The perception of marking reliability amongst teachers is a major problem for key stage 3 English. Unfortunately, however, the solution to this problem is far from obvious. It could, of course, be overcome by making English tests more like maths tests; that is, by dividing papers into much smaller components and by asking questions that have relatively unambiguous answers. Yet this could compromise the validity of the tests and would, most likely, lead to an even greater outcry from key stage 3 English teachers. On the other hand, problems of confidence amongst teachers appear to be less serious for English at key stage 2. Problems of confidence amongst teachers also appear to be less serious for English at GCSE and A/AS. The implication, then, is that there may, perhaps, be lessons to be learned from approaches taken to the marking of English in other assessment contexts.
Finally, when differences between EMAs were explored, the only significant ones were for the 'lost faith' statement and only for the KS3 English R3 sample: a smaller percentage of AQA-N respondents appeared to have lost faith in the review system. Whether differences in the processing of KS3 English review requests at AQA-N had made an impact was not unambiguously indicated; however, the fact that the Lead Chief Marker was based there (assisting the Lead Chief Marker for KS2 English in processing the requests) might suggest so.
Section 6 Professional perspectives on the processing of KS2 and KS3 review requests: strengths, weaknesses and scope for improvement (Study 5)

6.1 Introduction

The system for processing national curriculum review requests evolved gradually over the past half-decade without being subjected to substantial overhaul or evaluation. The present research project was therefore the first major opportunity for a range of stakeholders to step back from the yearly cycle and to consider the extent to which the review system functioned as effectively as possible. While Studies 3 and 4 gave schools the opportunity to express their views, Study 5 gave those directly responsible for the effective functioning of the system their chance to comment.

In contrast to the intention of Study 1, which was simply to identify the procedures that were being employed by the English EMAs, the intention of Study 5 was explicitly evaluative. The focus was to be upon identifying the strengths, weaknesses and scope for improvement within the system. Once again, it is important to stress that the guiding principle underlying the present project was to focus particularly upon scope for improvement. This inevitably meant that weaknesses would be over-represented in the final report. If the overall impression, from the presentation of Study 5, appears unduly negative this conclusion should be resisted. Instead, the report should be read constructively as indicative of the potential for improving a system that appears already to function effectively.

6.2 Methodology

The approach adopted was to identify a sample of key professionals within the review system and then to invite each one to take part in a face-to-face interview that would last around one-and-a-half hours. The interviews would focus on the dimensions of the review process that had been identified during Study 1 and would be explicitly evaluative.

The interviewees were selected so as to represent the following interests (within parentheses are the abbreviations by which the interviewees are referred to within the present report):

1. AQA-S External Marking Agency project managers (EMA);

2. AQA-N External Marking Agency project managers (EMA);

3. Edexcel External Marking Agency project managers (EMA);
4. Mathematics Lead Chief Marker (LCM);

5. Science Lead Chief Marker (LCM);

6. English Lead Chief Marker (LCM);

7. QCA External Marking project manager (QCA);

8. NDCA project manager (NDCA);

9. DfEE project manager (DfEE).

The sample was selected by the NFER project manager, although QCA advice was sought for the selection of representatives from QCA, NDCA and DfEE. All participants that were invited agreed to participate. At AQA-S and Edexcel both manager and deputy manager responsible for reviews took part in the interview. At AQA-N only the deputy responsible for reviews was interviewed. As he also happened to be the Lead Chief Marker for key stage 2 English, this meant that four of the six Lead Chief Markers were interviewed in all.

It was originally intended also to interview three Chief Markers. However, this raised the problem of which of the eighteen Chief Markers to invite: which EMAs, which key stages and which subjects? On further reflection, it was decided not to interview any of the Chief Markers but to invite all eighteen, and the remaining two Lead Chief Markers, to submit written responses by post. Written responses were received from one of the Lead Chief Markers and from nine of the Chief Markers. Overwhelmingly, it was the mathematicians that tended not to respond. All interviewees and respondents were asked to speak in a formal capacity as organisational representatives; however, a number of participants were also able to offer comments from other perspectives (i.e., as headteachers, GCSE examiners, etc.).

Finally, it was explained to EMA project managers that the NFER would welcome formal written submissions. In fact, the EMAs chose simply to respond through interviews.

Before the interviews were conducted a generic interview schedule was developed (see Appendix 6.1). This was structured around the key areas identified during Study 1. Within each area, a number of questions were posed to help stimulate discussion. Some of these questions were specific, although many were general. It was stressed that the interview schedule was intended only to stimulate discussion rather than to limit it. Participants were encouraged to address any issues that they wanted to raise; likewise, it
was explained that there was no expectation that they should comment on all aspects of the interview schedule.

While the interview schedule was designed primarily for the interviews it was also sent to each of the Chief Markers by post. Again it was stressed that the document was intended to stimulate, rather than to constrain, their responses. Admittedly, the written-response format was not ideal for collating the Chief Markers’ impressions of the system. Still, important issues were raised by them that would not have been identified without this consultation.

A number of the interviewees did not have direct responsibility for the day-to-day processing of review requests within EMAs (especially, DFEE, NDCA and QCA but also the Lead Chief Markers). Where appropriate, their interviews departed from the schedule to consider the specific role of the interviewee and their unique perspective upon the system. For all participants, the interviews began with a few introductory questions concerning their impressions of working relationships within the system. Participants were also invited to close the interviews by highlighting what they perceived to be the particular strengths and weaknesses of the system and what they considered to be the most pressing issues for improvement.

Interviews were conducted during the end of August and the beginning of September by the NFER project manager. These were audio-recorded and transcribed for analysis. Analysis involved coding each of the transcripts and selecting key passages to illustrate major emergent themes. A draft copy of the report was sent to all interviewees and respondents. They were encouraged to check all quotations to ensure that they had not been misrepresented. Where comments we received from participants, including suggestions for rewording quotations, these were integrated into the second draft.

6.3 Results

The presentation of results does not follow the generic interview schedule question-by-question. Instead, results are presented under a variety of sub-headings within the broad area headings that also structured the presentation of results for Study 1. In addition, the presentation begins with a section on the general functioning of the review system and ends with a section on overall impressions. The results are presented discursively and quotations are used liberally for illustrative purposes.

6.3.1 The general functioning of the review system

The interviews typically began with a number of questions that addressed interviewees’ perceptions of how the review system functioned at a general level. These touched upon
issues such as: whether communication was effective; whether working relationships were sufficiently well defined and understood; whether procedural requirements were explicit enough; etc. These issues are of particular importance, as the hierarchies and structures of national curriculum assessment are undoubtedly complex.

While the QCA is responsible for the external marking of national curriculum tests, this is undertaken under the guidance of the DfEE, through which it is ultimately accountable to the Education Secretary. Furthermore, the QCA devolves responsibility for external marking to the three independently functioning EMAs, who, in turn, appoint markers, on a part-time basis, primarily from the ranks of the teaching profession. Somewhere, in the middle of this structure are Lead Chief Markers who are technically independent of all participating organisations, but who are actually appointed by one of the three EMAs.

6.3.1.1 The review system in perspective

“This year, there has been immense good will to get it (the review system) right.”

NDCA

There was a general feeling across interviewees that – while there had been problems in the past – the general functioning of the external marking system had improved radically in recent years, with strong management from QCA and with good will from all parties involved.

“the whole review process … evolved a bit on the hoof … gradually we’ve moved towards a more formalised and structured system … and we’re getting pretty much there now.” LCM

“the project management has improved three hundred per cent over the last two years … the whole thing is going in the right direction … it’s a case of making sure everybody’s on board and doing the same thing” DfEE

“I think there’s been great progress there recently. I think QCA have really tightened that up with step-by-step procedures.” LCM

The general impression was that effective management frameworks were now in place and that the next step was to tighten up ‘loose ends’. In particular, although significant progress in co-ordinating the work of the EMAs was recognised by all, there was still perceived to be scope for improvement in this respect.

“the notion is that the Lead EMA is going to lead the other two, no longer in a sort of primus inter pares … awfully nice collegiate sort of fashion, but if EMA X

nfer
needs telling to do something, the notion is that the Lead EMA will tell it and then it will be done. Now that ain’t working because EMAs are frankly offshoots of exam boards and exam boards have traditionally worked in competition with each other.” LCM

While it was clear from the interviews that EMA representatives were committed to working together, there was still an important message to be drawn from the previous quotation. The concept of a Lead Agency needs to be specified more clearly and the nature of working relationships between agencies need to be agreed and formalised. There are various ways in which this could be achieved; one suggestion was a radical restructuring along subject responsibility lines.

“I’ve put forward the idea that each of the three agencies should take charge of one subject and run it nationally. To give you an example, all the marking of English ... should be ... directed from ... AQA Newcastle. Guildford, perhaps, could direct all the science; Edexcel all the maths.” LCM

6.3.1.2 Structures and communications between and within agencies

“We [the QCA] certainly will make rulings; we sign things off; we demand things; you know, we have a sort of power relationship with them [the EMAs]. But on the other hand, they’re the people who have the expertise, or a lot of it, and they are the people at the front line.” QCA

The participants were happy that roles and responsibilities within the review system were generally understood and that, on the whole, channels of communication functioned effectively.

“one of the good things this year has been the relationship that we’ve enjoyed, both on the EMA side and the NDC side, with QCA ... We’ve been able to get round the table and discuss ... issues in a professional way. People have been able to understand issues from other people’s points of view. The resultant procedures have met most needs.” NDCA

There was still a feeling, however – particularly amongst the EMAs – that scope for improving mutual understanding existed.

“if we had a bit more contact with the DfEE, because it gets a bit like Chinese Whispers ... however much they would like figures earlier and earlier and earlier – it would just be nice if they just understood some of the constraints ... just to understand how sometimes it’s not as straightforward. And equally for us to
understand from them exactly what they need, and what is flexible and what isn’t.”

EMA

The least satisfaction with management and communication structures appeared to come from the markers. Indeed, there was a feeling amongst some of the Lead and Chief Markers that their contributions were sometimes overlooked.

“QCA rarely take notice of the Chief Markers’ reports.” CM

“I was not involved in [the development of the Schools’ Guide] until it reached draft-print stage – and that upset all the chiefs as well … we can grant that this year was exceptional and all the timetables started late, but it ought to be that all the things to do with training have consultations with the people who train” LCM

It was especially evident from Lead Chief Marker comments that management structures had led to communication problems in the past. One LCM noted an incident where the rapid turn-around of a key document had been hindered by the formal approval process. This highlighted the need for all involved to take steps to ensure the smooth passage of to-be-approved documents through management structures; but it also drew attention to the particular communication problems facing LCMs. This is not simply that all communication between QCA and LCMs should be channelled through Lead Agencies, which adds an extra link in the chain; it is also that the chain splits in two at QCA, as communication may need to be directed either to the assessment division, or to the subject division, or to both. Moreover, in addition to the problem of filtered communication with QCA, the LCMs also face problems of filtered communication with their chiefs.

“If a message about English marking has to be got through the system, then the Lead Agency has to go through the process of protocol contacting the relevant Project Officers at the other EMAs who then contact their Chief Markers … if the whole thing was run from one place for each subject you wouldn’t have to do that.” LCM

6.3.1.3 The specification and documentation of expectations

In previous years, the EMAs provided review services to agreed specifications, but on their own terms. This culture has changed over the past few years and there is now an expectation that the same procedures should be applied in the same ways across EMAs. However, it became apparent during the interviews that clearer specification of expectations was required in certain areas and that explicit documentation of intended practice was necessary.
“there’s nothing written down … it’s just something that’s evolved over the years” DFEE

“I think inevitably some things have been lost there, because ultimately not everything is written down … things are carried around in the heads of people who’ve been doing the job for two or three years and who know the pattern of things.” LCM

“I don’t think [QCA] have set out sufficiently well what they expect each EMA to do.” EMA

QCA explained how it was hoped to remedy such problems through a new project management methodology (PRINCE2) which is structured around product descriptions.

“in that product description you describe: what the product is; what are deliverable; what its purpose is; how it’s made up; what its composition is; its format and presentation; what it actually looks like; where it’s derived from; quality criteria” QCA

Although the development of product descriptions required additional work of both QCA and the EMAs during 2000, there was evidence that it was having a constructive impact.

“product descriptions, managed well, can be something that the whole department takes part in; because it’s the people who actually stick the labels on who can tell you how to stick labels on … not those of us who say ‘stick labels on’.” EMA

One of the issues that caused significant problems for 2000 was reporting on the progress of reviews to QCA via Management Reports. It appeared that a contributory factor was the lack of explicit definition of terms.

“the definitions aren’t tight enough … we’ve reported something and the QCA have gone ‘this can’t be right’, because their interpretation of what they thought we would be reporting is … possibly not … our interpretation of what they were actually asking for on the form.” EMA

“The definition of a complete review; is it a complete review when the marks are complete? Or is it completed when it’s entered into the system? Or is it completed when it’s actually, you know, sent back to the school?” EMA

In light of problems like this, one EMA suggested that it might be useful to develop an EMA guide to reviews, to complement the Schools’ Guide.
6.3.1.4 The revised role of the Lead Chief Marker

In previous years, Lead Chief Markers had no formal national role in the co-ordination of review re-marking; they were appointed as Chief Markers within EMAs which is where their primary responsibilities remained. For 2000, LCMs were no longer to act as Chief Markers within EMAs but were to have a more independent role. Their *Duties and Terms of Appointment* for 2000 set out the following with regard to reviews (from the 2000-2002 Tender Specification, Annex D):

A10 To assist QCA and all EMAs in the planning of review meetings to ensure that all EMAs achieve the same quality of conduct of reviews. To be responsible for ensuring that consistency with regard to the application of the mark scheme to reviews is taken across all EMAs and in the quality of reports and responses provided for schools.

A11 Where appropriate, to select in good time, standardisation material for use by markers preparing to undertake reviews. In addition, the lead chief marker will be responsible for preparing standard responses for use in preparing reports to schools on the outcome of reviews and for working with the lead EMA to circulate these responses to all EMAs.

A12 To mark review requests as allocated by the lead EMA.

It became clear, from interviews and written responses, that these requirements were not well met. Indeed, neither of the mathematics LCMs had had any direct involvement in the review process at all during 2000.

“One of the key things about this interview, in fact, is to focus my attention on the fact that I’ve had nothing to do with reviews in England, other than to tell my Chief Markers that the only game in town is mark them properly ... I’m there to be consulted, it’s just that, I presume ... they don’t need me ... the fact that I’m not involved in the reviews isn’t a panic job. I mean, there’s good people doing it.”

LCM

As explained by one EMA, the role of the LCM seems to be interpreted slightly differently depending on the personality of the person who is undertaking it. There was clearly a great deal of uncertainty, amongst all involved, concerning exactly what was expected of LCMs.
"there have been tremendous inconsistencies across subject areas in terms of the support that we've received ... almost to the extent to which you actually feel that you can approach them and ask for advice and guidance." EMA

"I would have welcomed the other Lead Chief Markers indicating that they would like to come [to review panel meetings]. To some extent I feel that they need to initiate that." EMA

The issue of to whom the LCM is accountable is far from clear, as is the issue of whom the LCM had authority over. Being appointed by one EMA, exactly what authority does an LCM have over the other EMAs (i.e., do they have the authority to require all EMAs to follow a common practice)? Or, alternatively, does authority reside primarily with the Lead Agency that appointed them? Indeed, to what extent are LCMs under the authority of the agencies whom they are not appointed by (after all, each EMAs is ultimately responsible for the delivery of their own review service)?

Different LCMs took very different approaches. As indicated earlier, the mathematics LCMs were not directly involved with reviews. The science LCMs were involved to a greater extent and there was an effort to provide standard responses for other EMAs to facilitate consistency. The English LCMs had most involvement and both provided re-standardisation material. One LCM was also able to attend review panel meetings for all three EMAs.

"I'm always a little bit different because I'm ... the one full time Lead Chief Marker." LCM

"I've been [at AQA-N] preparing reports to go back to schools, and that's something I've done in the past because I've been a Chief Marker. Now, arguably, as Lead Chief Marker, there's no reason why Guildford or London shouldn't ask me to go and help them prepare some of their reports to schools." LCM

There was also some concern over whether LCMs should be appointed by individual EMAs, if the whole point on the new role is to be independent.

"the Lead Chief Marker appointment should be a QCA appointment, because being an Edexcel employee can compromise me very severely ... I think it's a charade anyway because, certainly this year ... Edexcel were told [by QCA] to appoint me." LCM

Concern was also expressed that the revised role of the Lead Chief Marker, as described in the Duties and Terms of Appointment, might actually impose an unrealistic burden.
"[the requirements upon LCMs to undertake a national monitoring role] are very demanding and ... I think they are impossible to meet. If they want to build in more time and pay more days work then I think ... there's room for development in it but ... I can't see how I can monitor the reports written by other EMAs to schools on the outcomes of reviews other than if ... they have to select a sample and send to me ... [but] then I haven't seen the material, I don't have those marks, I'm only ... looking at the quality of English in the reports ... or what am I doing?" EMA (also LCM)

It would be wrong to give the impression that the revised role was entirely ineffective, though. On the contrary, it helped to facilitate some important innovations, at least for science and English. What was clearly apparent, however, is that the role needs to be defined more explicitly.

"I felt the way in which LCM and CM divided responsibilities for review worked effectively and panel members responded positively to the role of each. This is the first time that there has been a clear sense of a national perspective." CM

6.3.2 Procedures for informing schools of their right to a review

The interview schedule distinguished between two aspects of information for schools: information on how to apply for a review; and information on what would happen during that review. The *Schools' Guide* is the principal method for communicating information to schools and it attempts to meet both of these needs.

6.3.2.1 Guidance for schools on completing review requests

"As a headteacher as well as Chief Marker I feel that schools have sufficient information both on the review process and how to complete review requests." CM

Mixed views were expressed on the quality of information presented in the *Schools' Guide*. While it was generally considered to be at least sufficient, it was frequently felt that improvements could be made.

"it could be a lot clearer. Perhaps not in process terms, but maybe in the layout of the actual booklet itself. By just, you know, there are trivial ways of doing it -- have a signpost down the margin, clearer bullet-points rather than the sort of - it tends to be whole paragraphs of blurb which you need to read about three or four times before you get the gist of the one line you ... actually need to action." DFEE
Furthermore, some concerns were raised that, regardless of the quality of information presented, it was not necessarily being received as intended, or read by the right people.

“As both a teacher and Chief Marker I believe the information to be adequate. Unfortunately many schools/KS3 co-ordinators do not read the information carefully.” CM

“I think there is a possibility that within a school the information is there but not necessarily with the person who is doing the particular job at the particular moment in time.” LCM

On the other hand, it was acknowledged that schools did appear to be getting better at requesting reviews, which would seem to suggest that the information was largely being read and understood.

Various improvements or innovations were suggested to get the message across more effectively. These included the use of examples within the Schools’ Guide, the provision of separate booklets and the provision of posters.

“it clearly is confusing to key stage 3 English about whether they submit the whole set of scripts or not, or a sample, or what type of request it is and so on; so that needs clarifying ... It would be of great help, I’m sure, to schools, if we gave them some examples of what to send in for what reasons.” EMA

“whether you need a separate booklet on reviews which is short and punchy ... you do this, this and this ... rather than sort of picking it out of some twenty page document.” DFE

“We always assume that the schools sit there and read paragraph-by-paragraph the Schools’ Guide – the chances are that they dip into it as and when they need to ... we were saying the other day ... can’t we do posters to inform the headteachers [of key deadlines, dates, etc.]?” EMA

6.3.2.2 The transparency of the system

It was frequently commented that the Schools’ Guide contained insufficient information on the review process itself.

“I don’t think there’s enough in the book about the nature of the review process ... in terms of saying who would conduct the review, giving the schools confidence that they would be very senior markers, that they would all be re-standardised, appropriately trained for this re-marking exercise.” EMA
“it could be made more apparent to them, the fact that ... we do treat this very seriously ... the fact that ... markers will not be re-appointed if we feel that there is a genuine question mark over their work.” EMA

The EMAs also felt that schools often misunderstood the relationship between themselves and the QCA, in terms of who owned the tests and in terms of who was responsible for laying down the procedural rules.

“there is still misunderstanding and this is one of the reasons why, in the Schools’ Guide this year, we have tried to lay out the fact that it is a team effort, it isn’t just exam boards doing this, it’s the QCA contracting other people do the other parts of the job.” EMA

“We’re working to the schedule as laid down by QCA ... But from the schools’ perspective ... it looks like we’re dictating it as Edexcel.” EMA

From a public-relations standpoint, DFEE commented that there was a need to sell the review system more positively as schools still did not see it as a safeguard or a service for them.

“All you need to do, really, is sell what the review’s about; why it’s there; the fact it is a safeguard; the fact that if you need to request a review it doesn’t mean the whole system has no credibility whatsoever. And, you know, it’s not making excuses for the markers, or the marking system or the mark schemes, but it is saying that this is there to prevent all that” DFEE

More technically, one CM had concerns over whether schools understood the concept of ‘acceptable tolerance’ and asked whether this concept should be explained more thoroughly. It was suggested that technical issues like these might be addressed through INSET.

“Give national INSET on mark schemes explaining decisions on interpretation of pupil answers – this is done by Edexcel but not other boards.” CM

Finally, two CMs suggested that transparency could be enhanced by publishing review data by EMA, rather than just presenting national figures.

6.3.3 Procedures for the administration of review requests

A variety of issues were addressed in the general administrative section. They focused primarily upon procedures for requesting reviews and for processing those reviews within EMA offices. The EMA officers were best placed to comment on the finer points of
administration while other interviewees and respondents contributed comments on general issues of procedure.

6.3.3.1 The process of requesting a review

"The quality of review applications is generally felt to have improved this year, I have to say." EMA

Although it was generally felt that schools were getting better at requesting reviews, it was noted that many schools still made errors. Numerous problems were identified, such as: forgetting to enclose deposit cheques; sending requests in late; failing to specify samples for R3 reviews; omitting the headteacher's signature; not explaining their requests fully enough; failing to send in all of the relevant scripts; sending in photocopied scripts; not sending mark sheet photocopies; not filling in forms correctly; sending in requests for pupils whose levels would not change anyway; etc. As one EMA explained, making sure that all problems are dealt with immediately introduces unnecessary delays into the system. On the other hand, not being able to sort out all problems immediately can lead to delays at later stages.

"I'm actually still waiting for schools to send in stuff now [08/09/00]; I mean there are three or four schools ... [for whom all of] their scripts need to be looked at. And I haven't got them. We've requested them." EMA

"a school with six classes collects in the scripts for five classes and forgets Mr Bloggs who was off sick that day or out on a trip ... and then you get a supplementary package. So it's little things like that." LCM

"So missing items is the real frustration; it's not large ... in terms of numbers but it's large in terms of stopping systems and progress." EMA

A number of interviewees, particularly DFEE and QCA, felt that the entire process of submitting a review was longwinded and in need of streamlining.

"Schools are under the impression that they have to sit down and literally mark or re-mark every script, which is nonsense and you shouldn't really be needing to do that. And then, once you've done it, identifying each script or each mark which was wrong, ring it, photocopy it, fill in three separate sort of forms, photocopy the forms from the [Schools' Guide] - you know there's got to be a way of making that whole system easier and, you know, cutting out the time actually involved in doing it." DFEE
The issue of whether or not the review process, as it currently stands, effectively expects teachers to re-mark each script is a moot point. There is currently no formal guidance on how teachers should check their pupils’ scripts. However, there clearly are many hoops for schools to jump through when requesting a review and these would benefit from being revisited.

6.3.3.2 The deposit

“I mean the whole reason, as far as I can remember, for having a deposit when you request reviews was to deter people from actually doing it when it first started. You know, maybe that’s an adverse side of things, maybe that should be taken away.” DFEE

It was not at all clear, from participants’ responses, exactly what the deposit was intended to function as. Most, though, felt that it served a mild deterrent function. As such, the rate appeared to figure less in comments than the simple fact of its presence.

“I don’t think we can abolish the notion … ‘it’s a free service, send in your scripts if you want’ … because we would get far too many then.” LCM

“There ought to be some tariff, just to make them think and, obviously, a reimbursement if everything goes through because you don’t want it to be a deterrent, only something to keep their minds on track really.” LCM

“we’re not really looking to penalise schools if the marking is incorrect … but we do need to stop them just throwing stuff back.” EMA

When the rate was discussed, it was generally felt not to be too high; if anything, it was felt to be too low.

“I mean that’s chicken feed really, considering the costs and the admin. and all the rest of it. I think it could go up.” LCM

One issue that did raise concern was the fact that deposit cheques were returned on the basis of whether any of the pupils in a school’s submission for a subject received an upgrade. It was felt that this might be disingenuous if it encouraged schools to submit scripts that should not have been submitted. Simply from a statistical perspective, the more scripts are submitted, the lower the chance of a school losing its deposit.

“I think it’s ridiculous where they put in thirty R2s – that’s £150 – and they only have to get one level change and they get the £150 back.” EMA
“You know we’ve got some science scripts upstairs and they’ve just been marked up ‘this will be worth a try’, ‘let’s try this one’ … ‘probably got one dead cert.’, ‘let’s put in three or four, we’ll get our money back.’” EMA

Finally, one EMA suggested that, to overcome certain deposit-related problems, it would be possible to invoice schools on completion of their review (but only if the school was unsuccessful).

6.3.3.3 Paper-based administration systems

The review system is, essentially, paper-based; schools return paper scripts, with paper mark sheets, and fill in the details of their requests on photocopied sheets. To keep track of the progress of these requests, the EMAs complete tracking forms and all items are ultimately filed for reference. In some senses, this means that substantial bureaucracy is involved in the processing of review requests. However, it was generally felt that the bureaucracy was both necessary and effective.

“this bureaucracy that we have is necessary for sake of efficiency given the number of requests that we have ... we need hard logs which follow the review requests wherever they go.” EMA

“We also have a paper-based administration system for R2s, for R3s: who’s got it; where is it going ... and I don’t think we’d ever depart from that because it’s very cosy.” EMA

Whether or not the system might be streamlined, though, is still open for discussion.

“It’s all paper-based system, you know, erm, IT doesn’t feature in it. It means that everything has got to be logged. It’s got to be logged when it comes in, it’s logged when it goes out to the markers, it’s logged when it comes back in again ... you wonder why we can’t streamline it in some way.” EMA

“We’re going to look at it and see whether we can come up with a single log sheet rather than ... a different log sheet for each type of review, so we are reviewing that.” EMA

6.3.3.4 IT-based administration systems

In parallel with their paper-based systems, all EMAs also logged the details of all requests onto databases. These were used to generate Management Reports for QCA and for the exchange of data with NDCA. Although the systems appeared to function reasonably
effectively, they were still relatively new and it was clear that not all problems had been resolved.

“certainly the IT ... we’re really in the dark ages here compared to what they’ve got [for GCSE/A/AS].” EMA

“There’s undoubtedly scope for improving the IT.” EMA

“Our ICT-based system; it can be more effective.” EMA

6.3.4 Procedures for selecting, training and standardising review markers

6.3.4.1 The selection of review markers

The interview schedule asked whether there were any problems in selecting review markers. Participants’ comments indicated that ensuring sufficient markers were available to process reviews on time did sometimes cause significant problems, particularly for English.

“markers, certainly key stage 3, move onto GCSE marking ... so that reduces the marking force. And, obviously, holidays. So ... getting enough markers to cope with review requests is a big management headache for EMAs and, generally speaking, they’re doing very well. But that can add in extra time, and it can mean that things are more difficult ... for them to manage.” QCA

One of the requirements upon EMAs is that review re-marking should only be carried out by grade A supervising markers. Clearly, this limits the pool of potential markers. But it also raises some more subtle problems, even down to how a grade A marker should be defined (past performance versus present performance).

“What we do is write to everybody in February ... and say ‘are you available?’ ... The fact that we had so many reviews this year did present a problem ... You can’t plan for those because you don’t know in February who’s going to be an A grade marker.” EMA

One EMA also explained that they did not invite markers who had reviews against them which, again, might lead to problems due to early selection. Another, on the other hand, was less sure about how to deal with such situations.
“Do you automatically eliminate that person ... it may not be an entirely genuine request ... we try to avoid that marker knowing [that they have a request against them].” EMA

A question mark was raised over whether only supervising markers should be appointed, particularly if the task does not involve supervision.

“you’ve got to have simply accurate marking ... I don’t see why, for review marking, we shouldn’t even have experienced ordinary markers, not team leaders or supervisors, who have consistently been graded A for their competence and accuracy in marking. Because we’re not asking them to do anything except mark these scripts immensely accurately.” LCM

One issue that had not necessarily led to problems in the recruitment of effective markers, but that, at the very least, might have caused resentment amongst those appointed, concerned marker fees. On the one hand, it was questioned whether markers, in general, were paid enough; on the other, it was suggested that markers in different subjects might deserve different fees.

“in Australia ... they’re being paid something like twenty pounds an hour for marking the scripts ... I think to attract a better quality of marker you’ve got to pay more money.” DFEE

“if the key stage 3 English markers do review marking on R3, they get paid one-and-a-half times the script fee. If they do marking for an R2 they get paid the script fee. Strictly, if they only mark one paper they only get half a script fee; yet they have to write a review, they have to write a report and they take time and markers literally loathe doing it ... it does seem very unfair that ... science markers can [mark quickly while] it can take 20-30 minutes sometimes to do one review [in English].” EMA

Finally, there was a suggestion that review marking should be a requirement for senior markers, rather than an option.

“The effective management of the marking process does not end when scripts are sent back to schools and this should be reflected in the contracts of senior markers.” CM
6.3.4.2 The necessity of formal re-standardisation

As part of the process of tightening review procedures for 2000, there was an implication that markers should be re-standardised. However, it was left unclear exactly how this re-standardisation ought to be conducted, who would ultimately have responsibility for the re-standardisation and the extent to which practice should be allowed to differ between subjects and EMAs. As a result, re-standardisation was not at all consistently handled.

The interview schedule questioned the extent to which re-standardisation was necessary at all. Responses varied considerably, but primarily along subject lines.

"I'm confident that I can defend the claim that there is no need for re-standardisation for reviews. I mean, it's business as usual." LCM (maths)

"Not at all, as I do them all." CM (maths)

"Not at all as only the Chief Marker or Deputy Chief Marker undertake reviews and they have worked as a team and do understand the application of the mark scheme." CM (science)

"I don't think we need that re-standardisation to the extent that English does." LCM (science)

"it is clear that re-standardisation in some form is necessary in English." CM (English)

"I do think it's important ... it is important as a check ... we are reminding ourselves of everything that's in the mark scheme ... a lot of people who mark key stage 3 also examine GCSE. If they've come from key stage 3, to GCSE, back to key stage 3, I think that ... emphasises the need for re-standardisation." LCM (English)

"In English, that is particularly important because English is so difficult to mark ... and in science it's a mid-way type of issue ... maths is virtually no problem – it's right or wrong." QCA

Many of the interviewees and respondents were keen to suggest an important distinction between formal and informal re-standardisation. This highlighted the role of the panel meeting. AQA-S. most notably, ran panel meetings for all subjects; all review markers for each subject gathered on specific days to begin processing requests as teams. This gave the opportunity for informal discussion of individual scripts when problems arose.
Moreover, for English, it also gave the opportunity to begin with an informal re-standardisation using pre-circulated scripts.

"The meeting in July is not seen by any of us as re-standardisation, but clarification of the rules of reviews, i.e., consistency of application of mark scheme.” CM

"Formal re-standardisation is not necessary ... Spending a morning in re-standardising together is more valuable than jumping through further pre-determined hoops.” CM

"[In previous years] review meetings, at which the samples were marked and decisions taken, were held in Edexcel premises in central London on Mondays, Wednesdays and Fridays during July and August until reviews were complete. Re-standardisation was unnecessary because samplers worked in pairs under the direct supervision of the Chief or Deputy Chief Marker.” CM

"I think we ought to have panels set in subjects ... you can re-standardise collectively, it’s not remote, it’s part of a discursive re-standardisation and it isn’t like the sampling where they send a bit of paper with their results on ... If it’s remote standardisation ... there’s less reason for it and less positive outcomes of it.” EMA

While panel meetings were generally viewed positively, issues were raised concerning exactly how they should be organised. One CM felt that re-standardisation scripts should be circulated in advance so that Sample Script Record Forms could be scrutinised before review markers were selected. One of the EMAs wondered whether it was necessary to re-standardise for review if markers had already been re-standardised for the Post-Completion Check (in KS3 English). One LCM felt that panel meetings were most successful when those that spread over a number of days were held on consecutive days – otherwise, it encouraged inconsistent attendance which would mean that some markers would need to be re-standardised remotely by telephone or by post. He emphasised that, for English at least, the face-to-face interaction of the panel meeting was all-important. Finally, while the cost-effectiveness of panel meetings was queried on a number of occasions, it was generally considered that the benefits outweighed the costs.

"Well, the whole thing’s budget-constrained, quite severely; so EMAs will not be wanting to pull people in from afar. If they can handle it all at a distance, like Edexcel does, they’ll do it because it’s cheaper ... one could argue that the model
of centralising it might well have a bit of extra cost but I would argue the point that it was seriously worth it.” LCM

“I think panels are more cost effective than remote marking ‘cos you pay them a cheaper rate – but that is not the reason; they’re more cost effective in terms of also the quality, the standard that you can guarantee.” EMA

Although the necessity of re-standardisation was generally discussed from a technical point of view, it was also realised that there was an important dimension of accountability.

“I think it’s very important to re-standardise, to be able to demonstrate to schools ... to offer ... a degree of reassurance to them that we’re doing actually everything possible to ensure any re-marking ... is done as accurately and consistently as it can be humanly be made to be.” LCM

6.3.4.2 Review marker training

Perhaps the most significant oversight in the system for 2000 was the lack of formal training of review markers to undertake those aspects of the process that differ from initial marking. To some extent this had been addressed for KS3 English, where markers were given training on the decision-making process for R3 samples. However, more generally, there was a distinct lack of training, across subjects and EMAs, on how to write review reports.

“even if they may be a grade A marker, they may not be a grade A report writer.” EMA

“you get the difference between the people who write ‘this is not in accordance to the mark scheme’ and those that write you a small essay.” EMA

“the ... area in which one ought to have a consultation ... is unification of reports. What do they say in reports? What should we say in reports? I think it’s common sense but it might be just as well for us all to agree what the common sense is.” LCM

“we’re then trying to put together a report ... on [the behalf of markers that] doesn’t necessarily highlight the key questions that the school would have wanted answered, because the marker hasn’t been quite clear on what they should be reporting.” EMA
"Well, the [KS3 English] markers send in their written comments, and the best of these you can virtually lift. The worst of them you have to do a lot with. Most of them write more than is needed. They're paid ... £29 to write this report, so they tend to give you what they think is £29 worth of writing. And you don't want detailed specific comments on half a dozen characters, you want ... [for example] 'the marking of question four was generally accurate, but the marking of question three was shambolic'. Perhaps one or two brief examples" LCM

6.3.5 Procedures for undertaking review marking

6.3.5.1 Time pressures

It was generally accepted that, although time was very tight, the turn-around of review requests was usually manageable within the specified time limits. There were a few suggestions for how the process might be speeded up; for instance, one CM suggested that EMAs might send scripts to review markers as soon as they received them from schools (rather than batching them up and sending them together). On the other hand, as an LCM commented, despatching scripts piecemeal, as they arrive at the EMA, means a large administrative burden which might actually increase the time taken to process reviews. Of course, the issue of despatching scripts interacts with the issue of whether a panel meeting is scheduled. If an EMA runs panel meetings then no R2 or R3 reviews will be considered before these are begun.

An additional point, raised by an EMA, was that few review markers who were also teachers would be prepared to undertake re-marking until after schools had broken up. The end of term is simply too busy a time to be undertaking additional work. The different profiles of marker professions in different EMAs might, therefore, affect the speed with which re-marking could commence.

6.3.5.2 The bureaucracy of re-marking

The question of unnecessary bureaucracy was raised under a number of headings within the interview schedule. The considerable burden of filling forms was raised by markers in English, maths and science.

"In 2000 there was an inordinate amount of form-filling to indicate, for example, changes in levels of individual pupils. It was not clear why this was necessary and its usefulness was doubtful. There was some expression of frustration amongst the team about the time taken to complete these details." CM
"There is duplication of paperwork. The form-filling relating to the review marking takes considerably longer than the actual review marking itself." CM

"the amount of paperwork that the reviewer actually has to manhandle is bonkers really ... I mean, it’s complex, especially if you want to ... turn down a thing. It’s easier to uphold an appeal, from a reviewer’s point of view; just tick the box and ‘thank you very much, upheld’. Whereas you’ve got to do a bit of writing if you want to turn the appeal down. I don’t know what impact that’s going to have on the number of reviews upheld and turned down but certainly there’s a disparity in the workload there.” LCM

6.3.5.3 Standard comment banks

The processing of review requests was facilitated significantly for science through the provision of standard comment banks. The comment bank was a list of responses to common queries. Each comment was identified by a discrete code; markers simply recorded the code as appropriate and the EMA automatically converted the code to a standard response when finalising the report to schools.

"My list of coded responses save me and the review markers a great deal of time; I hope it works in the office ... the use of coded responses ensures consistency.” CM

"I have wondered whether some of these standard statements have been a bit brief, but I have not had any schools come back and say ‘you know, we don’t know what’s going on’ so they’re short and sweet and to the point.” EMA

AQA-S, the Lead Agency for science and developer of the idea, has successfully used the standard comment bank approach for a number of years, particularly at KS2. However, it has not been received positively by all science markers.

"each request is going to be different, with the school placing one interpretation on the answer and the ... review marker another.” CM

Likewise, the standard response format has not been adopted by any of the EMAs for either mathematics or for English. This issue will be returned to later.

6.3.5.4 Procedures for monitoring reviews

As for re-standardisation, procedures for monitoring reviews were not well specified in advance and, in general, it appeared that review marking was not monitored in any formal sense.
“There really isn’t [any monitoring of R2 or R3 reviews]. It is something one would like to do ... last year we actually managed to do some sampling on R3s, but this year with 171 reviews it’s just not been possible.” EMA

However, there was a general feeling, particularly amongst Lead Chief Markers, that review marking ought to be monitored and, moreover, that this ought perhaps to be the responsibility of the LCM.

“Should I have demanded that I see a sample of reviews in order that I have the quality assurance check personally of my own chiefs? ... it’s implicit, if the Lead Chief Marker is supposed to be assured that the review process is correct.” LCM

“There isn’t anything formal at the moment and one could argue that there should be. That would add a cost and delay in terms of time. One of the things we’re looking at ... is a sampling of live marking ... and you could envisage that with reviews.” LCM

“There is not an explicit system for monitoring in the sense that a school of 150 scripts has been re-marked and so 10 per cent, 15 scripts, are seen by the Chief Marker. That may happen, but it’s not a requirement. It depends on who’s around and available and doing what at any particular stage ... I suspect, despite the haphazard nature, there’s a fair amount of monitoring ... that does go on. But it could be formalised and possibly should be formalised. You’re getting the reassurance to schools to know that they’ve been sampled before and after, as it were.” LCM

6.3.5.4 Shortcomings of the review model

While specific review procedures may have been differently interpreted in the past, the underlying model has remained consistent since its inception. A number of participants took the opportunity provided by Study 5 to comment on this underlying model and to question whether it was necessarily entirely appropriate. The first issue raised was that the basic R2 model revolves around re-marking only specified questions rather than the entire work of a candidate.

“I would like to know why the full A-Z is not done. Is there some political thing that they don’t want schools to feel that, ‘... he’ll mark everything, so you’ll lose that one’... but that’s fair isn’t it?” LCM

“At present, only the item queried by the school is remarked. This means that often on the same script – even on the same page – there is another marking error
that is ignored. Very often reviews relate to poor markers, who make mistakes in both positive and negative directions. It would be ‘fair’ to give the pupil the correct level; this could be done by re-marking the whole paper if the point raised by the school is accepted.” LCM

Secondly, concern was expressed that, when a review identifies a dubious marker, their remaining allocation is not called in for re-marking.

“we’re only ever capturing the tip of the iceberg of the problem, if you like, because if you’re a marker and your scripts need re-marking, we only re-mark them if the schools have sent them in for a review ... You should really re-call the full allocation of all the scripts that they’ve had ... But we’re talking of a lot of expense and a very time-consuming process.” ... “also you would need to double the number of markers to actually see it through.” EMA

“There are occasions, and I have done it, where we have had a marker who has been quite off-the-wall and we’ve actually phoned up the other schools in our allocation and said ‘would you like to put in a review, do you have concerns, we will accept it?’ But you normally don’t do that” EMA

Thirdly, NDCA raised an important issue, concerning the way in which performance data are now being used, that might well have implications for the way in which reviews should be conducted; the issue being that Ofsted appears now to require secondary schools to demonstrate a two-third increase in pupils’ levels between Year 6 and Year 9.

“What we’ve found is that one of the uses of national data collection is to go to Ofsted to become the Pandas that go into schools ... It’s come to my attention within the past few weeks that Local Education Authorities are taking the mark ranges of the levels, for the KS2 tests, and dividing them into three ... So that they can produce, for their own purposes, sort of third-levels ... The question that’s posed in my mind is should we be going back and re-visiting the decision that only those ... mark changes that could affect the level should be the subject of a review? ... perhaps we should be ... giving better data at what are agreed sub-level points” NDCA

A fourth concern, that was raised by a number of participants, was the apparently unprincipled ruling that group reviews should only be permitted in KS3 English.

“why can’t you have a group [review for] key stage 2 English? ... there have been a number of schools who have, basically, sent the things back saying that they are
not prepared to fill in R2 forms for every one of them because they are very unhappy with the whole cohort.” EMA

Perhaps the most general, but also the most searching, question asked whether the rationale of the review service was sufficiently well defined.

“I think, perhaps, what we haven’t yet decided is what the number one priority to get out of it is. Is it satisfied schools? Is it data in on time? Is it things back to schools?” EMA

6.3.5.4 Miscellaneous review issues

A number of more specific issues were raised in relation to procedures for the processing of review requests. These tended to highlight issues where decisions were needed to facilitate consistent application of best practice.

A particular area of concern was the response that should be given to schools when an R3 review has been rejected, but the marking of the sample has revealed a small number of level changes for individual pupils. One EMA, for instance, was keen to know whether the 15 scripts ought really to be treated as a genuine sample — if they were, then the decision to take no action on the batch as a whole would imply no action should be taken on any of the 15 scripts. It was also felt that there was still some ambiguity over the R3 decision itself; while the training documentation was clearly appreciated and found useful there might still be some way to go to objectify the task further. Furthermore, the question remains as to who should be involved in, and ultimately responsible for taking, the R3 sample decision.

Finally, a few participants commented that there might be scope for a greater integration of statistical information into the review process.

“A particularly useful tool for review is the performance data for schools covering the past three years and indicating performances of cohort-on-cohort. This gives a clear feel as to what the marking outcomes, in level 6+ percentage terms, should be.” CM

6.3.6 Procedures for reporting outcomes of reviews to schools and the NDCA

6.3.6.1 Reporting review outcomes to schools

It became apparent, from Study 1, that a large amount of time and effort was put into the preparation of outcome reports to schools; firstly, by the review markers who draft
responses to schools’ queries; and, secondly, by the EMA administrative staff and managers who convert these drafts into letters and reports for schools. As noted earlier, far more time and effort is devoted to reporting upon requests that are rejected.

“This is where the pressure comes in. If the schools want the results, we can get them back quite quickly. If the schools want the reasons for the results, that does take longer ... to give you an example ... they go into boxes and then they ... all go into standard files in the PC and then you pull them up and you proof-read them. I reckon I can do a box full in a day if I’ve got a fair wind behind me, and not really having to do anything except do it ... I did five boxes full before I went on holiday, and that was just for 2 English and that’s how slowly you can proof-read to do the job properly.” EMA

“if you’ve got a standard letter that you’re going to insert paragraphs in, and it’s someone else’s job to identify the nub of the professional content of the response from the review marker’s report – [LCM] thinks it’s about 20 minutes or 25 minutes to do that so that’s his amount of time – and then that has to ... come through me and ... then it goes to the word processor and she just pops it into the letter ... I think it’s an appropriate amount of time and ... I’m not so sure that we ought not to write a fully unique report for the school” EMA

On the interview schedule, the first question on reporting review outcomes asked whether the time and effort required was justified. The majority of participants were happy to affirm that it was.

“In today’s educational climate few are ambivalent about results. It is understandable that they wish to attain a higher position in the league tables. Therefore the time spent on reviews is without doubt justifiable. Time spent on rejection reports is justified when schools have been specific in their requests, making reference to the mark scheme.” CM

“If we want students and schools to improve then the report back to schools – especially when I reject the request – is important. Unfortunately, I can’t give even longer rationales.” CM

“I put great value, and I know my chiefs put great value, in their report; because it is a personal request from a school, it means a great deal to the person being reviewed and we treat it as such and so ... particularly if we do not uphold it, we take great pains to explain to them” LCM
“Yes, totally justifiable. I think there are several benefits to the review process and one of them is educating – and I don’t mean this in a condescending fashion – but educating people in schools about science, about the tests, about mark schemes, about rigorous marking, and so on.” LCM

“There’s a sort of principle ... that has existed from the beginning of National Curriculum ... that this should be helpful to the teachers and to children in making progress, because it’s an assessment and it’s not a qualification. So, if you could feed back and say ‘this child didn’t do better on this question, because of x, y and z’ then that should be helpful to the teacher in future, knowing how to teach them better and how to prepare the children better to reach the standards that are expected at these points.” QCA

6.3.6.2 Standard outcome passages

Participants were invited to offer their thoughts on the use of standard outcome passages in reports to schools. As discussed earlier, these passages were at the heart of reports for science at AQA-S, where markers’ coded responses were converted directly into passages for schools. However, there was also evidence that a ‘standard passage’ approach could be adopted for English as well. For English, though, the use of standard passages was clearly post hoc; indeed, one attempt to collate standard passages from EMAs in advance failed completely.

“[Standard outcome passages for science make] it very quick for us to produce the review requests.” EMA (science)

“because I work on a computer, there is a certain amount of standard format writing, in that by the time I’ve done twenty of these I’ve got so many ... key phrases or sentences ... it gets into a bit of a cut-and-paste job. I wouldn’t want to go entirely down the road of ‘a hundred statements and you mix-and-match any twenty five’ sort of thing. But it’s a little bit like that; and it does speed you up.” LCM (English)

Less evidence of the use of standard passages for mathematics was noted. Moreover, a number of interviewees and respondents felt that the use of standard passages could risk being impersonal or giving insufficient feedback.

“a sentence is what’s required, because if they’ve gone as far as the review without knowing how to do it, then you’ve got to try and explain to them, yet again, what it is you’re trying to say to them, so I couldn’t do that in a tick box.” LCM
"In our experience, the use of these would significantly reduce ‘customer satisfaction’. Schools expect not only careful consideration of their request but also a personal and reasonably detailed answer.” CM

Finally, there was a suggestion that there was scope for speeding up the process of report writing by markers with the provision of word-processed templates.

“A standardised report form could be given to all review markers on disk, enabling the use of word-processing and speeding up the process of re-marking.” CM

6.3.6.3 Service turn-around times and reporting schedules

It was generally felt that service turn-around times were sufficiently well defined. It was felt that tighter deadlines would compromise the process.

“The turn round times are tight enough. If they were tighter then more markers would have to be involved ... I might lose the consistency I require.” CM

However, when asked whether outcomes should be reported to schools during the vacation there was a mixed response. This tended to be divided between those who believed that most schools were occupied during the vacation and those who believed that they were largely vacant. A number of participants suggested that there would be no reason why not to send outcome letters to schools as soon as reviews had been completed, although the return of scripts would have to wait until the autumn term.

6.3.6.4 Reporting review outcomes to the NDCA

Comments on the reporting of review outcomes to the NDCA came only from the EMAs and the NDCA. The EMAs, on the whole, were very happy with the system. The only specific concern raised was that replacement mark sheets from the NDCA did not always arrive as quickly as they might. However, one question did arise as to whether there might be unnecessary duplication in the information provided by EMAs.

“electronically letting NDCA know the changes of levels, through the CR3 programme, why do we need to send them top copies of the replacement mark sheets as well when they’ve already got the information?” EMA

The NDCA was equally happy with the way in which the system had functioned for 2000, and it was noted how co-operative the EMAs had been.

“I’ve always been struck by how hard-working the EMAs have been to give me what I need.” NDCA
Although happy that the system was functioning effectively, NDCA highlighted a number of areas where improvements might be made. The first concerned the format in which data ought to be received.

“We are concerned about the security of emails, even though the data is in a coded form. We would prefer a disk or the use of an ISDN link. Our fear is that if something went wrong, we would be blamed.” NDCA

The second point concerned how the process might be made more robust through a “belt, braces and bits of string” approach to data transfer.

“I think our biggest risk (in review data transfer) is that messages sent by the EMAS will not be received. What we have asked the EMAS to do this year is to number each review line as they send it. When we are loading a new batch of review changes from an EMA, the QA check is to ensure that there is no break in the line numbers which might suggest that a batch has been missed.” NDCA

A third point was not an issue of particular relevance for 2000, but was one that might become relevant in future years as more importance becomes attached to KS3 results. This would be particularly problematic for R3 reviews which tend to be the most likely to drag on into September.

“I close the KS2 database ... towards the end of August ... and then I don’t do any more work on it. And a KS2 review that got hung over at that point — not that I ever recall one — would be quite a serious issue. It would indeed be holding me up. KS3, nobody seems to bother. We don’t have school performance tables in KS3 yet, so it doesn’t matter if ... review results arrive in October.” NDCA

Finally, NDCA raised the issue of whether there should be a final post-review results confirmation mail-out to schools from NDCA.

6.3.7 Procedures for dealing with complaints against, or concerns with, the processing of review requests

“There is still a tendency for some schools to lash everything off to QCA or lash off a letter to DfEE, their MP, God knows who, and I think it’s important to emphasise to schools that other agencies don’t intervene ... Sometimes QCA will feel that, because they’ve been written to, they have to engage with it ... the only way to engage with it is a standard response of: we note your request; you should follow these procedures; if you have a further complaint at the end of that set of procedures then you may come back to us.” LCM
Structures for the management of the external marking system are complex and it may be understandable, to some extent, if schools are not entirely clear to whom they should complain when they feel they have been wronged. On the other hand, the Schools' Guide is very explicit in stating that requests for review will be dealt with by the EMA responsible for initial marking. Indeed, not only is their EMA the only channel through which a school may complain, but there is no formal appeal process, should a school feel that their EMA has not treated them fairly.

Participants were asked whether they believed that schools should have an opportunity to appeal against an EMA's decision. Views on this matter were clearly mixed. On the positive side, it was felt that the appeal was not just a basic human right, but also a mechanism for ensuring public confidence in the system.

"Everybody has the right to debate with you whether you’re right or not. They should pay for it, because if you don’t pay for it you’ll get some who may abuse the system ... but I think they should have the right to re-appeal and then it goes to a person who wasn’t the first one ... It’s got to stop sometime hasn’t it, but they certainly should have at least one further stab at it. We’re all human beings and we can all make mistakes ... it’s part of your transparency and guarantee to them.” LCM

"it’s a bit like an ombudsman role, isn’t it, if the school isn’t getting acceptable answers to its queries from the EMA then it ought to have recourse to a higher authority. So QCA is the obvious answer there.” LCM

On the negative side, it was felt that introducing an appeal process might be transmitting the wrong message about the system and that it would not be possible to establish defensible grounds for mounting an appeal.

"I think, if you suddenly started up with an appeals procedure, it’s going to give entirely the wrong message – and one that’s totally contrary to the message you have been giving so far really.” DFEE

"I don’t think they ought to have it, actually, if we’re trying to reiterate certain key things to schools. The more you say ‘well, if you’re still not satisfied, you could do so and so’. On what grounds would they have the right to? I mean what could they say which would persuade someone this is a valid response other than ‘oh, they’re still not satisfied, their profile isn’t right’?” EMA

Between these extremes was the suggestion that appeals should be allowed, but only in relation to whether procedures had been followed and not in relation to the re-mark itself.
“I am entirely happy with them registering a complaint against the way in which their review has been dealt with, but I don’t think they should have any recourse to a further review of the marks themselves.” LCM

6.3.8 Procedures for auditing, monitoring, or evaluating the effectiveness of the review process and for determining ‘customer satisfaction’

6.3.8.1 Auditing, monitoring and evaluating the review system

When asked whether the review system incorporated sufficient auditing, monitoring and evaluation exercises, the typical response was that it did not.

“No, there aren’t sufficient internal or external audits carried out and one of the reasons ... is that ... it’s an annual cycle ... and you’re just finishing off reviews and reporting on reviews and then it starts again ... whereas, if you had a period of time when you could actually ... sit down and reflect on it. But that’s an excuse because there ought to be time built in for more auditing and monitoring and evaluation ... I actually think we ought to do more peer monitoring.” EMA

On the other hand, many participants said that they would actively welcome the opportunity to improve this aspect of the system. To some extent, this would require resourcing that was not budgeted for within the external marking contract.

“I did make a strong point when this new contract came out ... that they didn’t ... have a budget for monitoring and review, and research and development, anything like that.” EMA

There was also some feeling that the system itself conspired against evaluation, for example, the one year contract for LCMs. One LCM suggested that this might be why key review data, such as which questions led to the most review requests, were not routinely fed back to him.

6.3.8.2 Sharing best practice

Again, although few formal mechanisms existed for sharing best practice between EMAs, participants tended, on the whole, to be in favour of closer collaboration.

“I think there should be the chance for more meetings between the EMA staff themselves ... that may be something that will come if we’re not so involved in product development.” EMA

177
"I welcome the opportunity to be fed with allegedly good practice – and I say
alleged because it sometimes doesn’t work for different subjects – but if I know
what QCA think is good practice and I know what AQA and Edexcel think is good
practice; I’d like an opportunity to have a look at it and then discuss with them:
would it work with us [KS2 mathematics]; and is it appropriate for us.” LCM

It was also commented that this mechanism would need to be extended to embrace Chief
and Deputy Markers and not simply project managers.

"A national review service should clearly operate according to ‘best practice’ and
a formal mechanism for establishing this would be welcome. This would need to
be informed by Chief and Deputy Chief Markers as well as by EMA staff.” CM

6.3.8.3 Customer satisfaction

When guided to consider whether the review system pays enough attention to customer
satisfaction, many participants again explained that more could probably be done. It was
often commented that customer satisfaction is a particularly thorny issue for the review
service because, if a school gets a request turned down, then it is unlikely to be satisfied
however well the service has been delivered.

"the final threat to a school or a group of schools who make a lot of fuss, is that
I’ll go and talk to them, I’ll pay a visit. And I do that from time-to-time ... in the
end, when there’s been a lot of huffing and puffing. And it may be that if ... QCA
goes down the route of setting up some INSET, then I think that would be another
area where we can work towards customer satisfaction.” LCM

QCA noted that one aspect of QCA’s annual evaluation of external marking is a
questionnaire to schools that does deal with schools’ attitudes towards the review process.

6.3.9 Overall impressions of the system

A number of issues that participants raised were broad and held implications for the
system as a whole rather than for individual component processes. The following section
begins by presenting commentaries on three general issues: implications for initial
marking; procedural flexibility; and the structural independence of review services. It
finishes by presenting issues that interviewees raised as particular strengths, particular
weaknesses and areas that they identified as offering the most immediate scope for
improvement. These summary comments arose at the end of the interviews in response to
an invitation to summarise impressions under these three headings.
6.3.9.1 Implications for initial marking

While the focus of interviews was explicitly upon review services, rather than upon external marking per se, comments on initial marking inevitably arose. These typically related ways in which the number of reviews might be reduced by introducing more effective procedures earlier in the system.

"... expenses incurred in re-marking (PCC included) could be reduced by more effective and rigorous sampling. As supervisory markers, we must either sample all schools within a marker's allocation or have sample sizes which reflect the size of the schools being marked." CM

"Well, I think that there probably are too many R1s from a quality point of view ... I suppose you could argue that doing 700-odd R1s is a lot better than actually ... going over every single mark sheet as a separate stage in the process before they go originally back to centres. But I suppose that the advantage would be that the schools would have the correct accurate results ... before the Summer break ... and they probably would have more confidence in the process as well." EMA

"for key stage 3 English we do not borderline at the N-3 borderline ... but we have seen an awful lot of reviews which have gone from N to 3 because – although this paper is not designed for these pupils – schools are putting in all their pupils; so I would definitely make a point that that should be included in the borderline check as well." EMA (manager 1) ... "I would endorse the borderlining at the N-3 and the 3-4 threshold ... but if QCA do want to go down that route then they probably will have to look at the fees paid to markers" EMA (manager 2)

"I would look at ways you can improve the quality control further back down the line, and the easiest way to do that – or the most effective way of doing that – is, first of all, to move the test reporting dates. Because the test reporting date, being so close to the test ... being taken, is the biggest stumbling block to improvements in the quality of marking." QCA

6.3.9.2 Procedural flexibility

The issue of procedural flexibility was of considerable significance in relation to the review system during 2000, as QCA had taken steps to minimise procedural variation between EMAs. In particular, it had targeted how EMAs should respond when schools do not follow the procedures for requesting reviews, as laid down in the Schools' Guide. QCA took the approach that consistency between EMAs would best be achieved if none of the EMAs gave schools much, if any, leeway.
"We’ve moved from a culture at ... QCA, where flexibility has been a first recourse; you know, schools haven’t managed to get their reviews by the set date, it’s a week late, the EMA says ‘what shall we do about it’, our response has been ‘oh, go on then’ ... This year we’ve said we’re not having any of that ... What we know is that it leads to inconsistency; it leads to people thinking that all they’ve got to do is shout louder and louder and louder and pull bigger and bigger noises in to support their case ... we’re clearer about the process in the Schools’ Guide, we’re clearer about what it’s designed to do, we’re clearer about what schools have got to do and what they’ll get from their EMA if they do it properly. We’re also clearer about saying ‘these are the rules and you have to play by them’.” QCA

While this was generally accepted by the EMAs, there was an underlying feeling that flexibility should not be outlawed entirely as it was essential to the effective running of the system.

“You’ve got to have flexibility at the end because there will always be a school with a special need or a special case or a special situation. An obvious one, in this region, is ... a little group of schools ... who break up for their summer holidays at the end of June and go back in the middle of August. And so sending in reviews for the normal time for them is barely manageable. So we’ve got to be flexible to accommodate whatever number of schools that may be. But you can’t have total flexibility; if you have total flexibility you haven’t got a system.” LCM

“rigour and consistency first and then the flexibility depending on individual circumstances” EMA

“we took the decision that we were actually going to get everything onto our systems so that we could get it out to review markers ... [then] in the holidays we faxed or wrote to all the schools and said ‘can you sign this and send it back to us’ and actually that worked quite well ... otherwise you could have actually stopped the review and said ‘oh well, we haven’t got it, we’ll send it back’ ... We have taken the decision that ... we’ll actually carry on and do it and get it afterwards. Now that may not actually be quite as we should have done, but what concerns me was ... that if we ... say, on a clerical check, we turn it down, that means that result is wrong. And I’m a little bit concerned; and I haven’t had breathing space to resolve it in my own mind.” EMA
6.3.9.3 The structural independence of review services

An issue that related to whether there should be a formal appeal mechanism was whether the present structures for processing review requests were sufficiently independent. This question was specifically posed to participants on the interview schedule. Although, many responded that the system was sufficiently independent, a number raised concerns with the public perception.

"there must be a perception that if those re-marking are effectively EMA employees then it is in their interests to uphold the original marking in financial terms." CM

"the schools are always going to be dubious about what’s actually happening – especially if it’s the EMA. I mean, if they got it wrong once, why should they get it right the second time?" DFE

A number of possible ways around this problem were suggested. The first was that the Lead Agency for each subject could be responsible for all review requests. This would inevitably facilitate consistency and more effective communication between review markers.

"the other idea that ... we really like ... is that the lead agency should mark all the reviews for that subject, which [because of the differential workload] would be pretty whizzy for the lead agency for maths, and not too bad for the lead agency for science, but a bit hard for our colleagues in Newcastle." EMA

"I do think that this notion of moving it to one EMA would virtually solve all of that independence stuff." LCM

A second possible solution was that reviews might be taken on by a national body, that was entirely independent of the EMAs, but that still appointed their senior markers.

"you could have a panel sitting for 3 or 4 days ... now whether that would help schools if they thought this was a national re-mark panel. But it ought to apply to all subjects and both key stages and I don’t see why that shouldn’t happen – taking it out of the regions and making a national re-mark project." EMA

A third suggestion was that the entire re-marking project could be taken on by a national body, that was entirely independent of the EMAs, and that appointed an entirely new set of markers. This might not only ensure independence, it might also help to speed up the processing of review requests.
“Do you need a professional body of people to re-mark that maybe ... aren’t teachers? ... that are going to be on-call all the time, that are going to be able to maintain consistency because they’re ... operating a set process.” DFEE

“Independent markers would be ideal; but presumably they would not be trained in the same way as original markers and ... review marking would not be consistent with prime marking.” LCM

The main problem with using an entirely new panel of re-markers has just been highlighted. But it is worth also highlighting the more subtle concern that review marking actually serves an important evaluation function and that might be lost if the present arrangements were changed.

“I would not want to do reviews on a national basis as review marking provides my last line of quality control on markers.” CM

“The review process is valuable as a check on any ... markers who submit ‘manicured’ Final Samples.” LCM

6.3.9.4 Particular strengths of the system

As indicated earlier, interviewees were asked to close by highlighting the strengths, weaknesses and scope for improvement that they believed to be most significant. These comments really speak for themselves.

“Well I think its main strength is ... it does go a long way to stabilising the credibility of the whole system ... the fact that you’re saying that we’re prepared to look at these things; you know, ‘if you think there’s an issue, a problem, then let’s look at it’; and ... we’re quite willing to look at it and re-mark it and admit if we’re wrong.” DFEE

“that there is a mechanism by which inconsistent or poor marking can be addressed, which is open, very open to schools ... because of the return of the scripts ... they have a mark scheme and they have a mark, so it’s totally transparent.” QCA

“you’ve got your best markers in ... So, from that point of view, I think you can guarantee quality ... everyone works very hard to make sure it works and down to the temporary staff there has been a great deal of conscientiousness” EMA
"I think the first key strength, which is the key strength of the whole thing, is the quality and commitment of the review markers and their commitment to getting it right." EMA

"Well I think the strength is the quality of the re-marking carried out by the re-markers — the dedication to that job." EMA

"we’re using senior people who’ve been involved in the development of the training materials, the development of the training programme, and who have, as I already said, imbibed all the issues. They are good scientists; they have vast experience in terms of teaching kids. So the personnel side of it is a major, major strength.” LCM

"I think the strengths, as I’ve indicated, lie in people working together and working as teams.” LCM

"This year I have a strong sense that all the people involved (in reviews) this year, have a clear understanding of what they are expected to do, and why. The increasing IT resources that we all have mean those systems are continuing to develop. It is not like a bureaucratic system in which people have forgotten why they are doing things.” NDCA

6.3.9.5 Particular weaknesses of the system

Once again, the major weaknesses highlighted speak for themselves.

"you’ve got to be very careful about how you handle that whole presentational issue, which is possibly a weakness because I don’t think we handle that well enough ... perhaps we should be a lot more up-front. And ... any changes that come out of this review that you’re doing, there should be big launches as to why it’s happening ... press releases, articles in the TES ... and make a big thing of it.” DFEE

"I think the biggest inequity is that there’s no court of appeal. I think that is something, I think there will need to be in future ... it will be an extra cost, but I think really there ought to be some sort of appeal.” QCA

"the schools are very rushed, so they’re pretty cheesed-off frankly; we’re very rushed because we’re getting reviews in but we’re still sorting out scripts that are not back. The timing is a weakness but I think there’s nothing we can really do about that ... the weakness maybe that we’re not, in the end, getting at everything
which we should be getting at. Because ... it's only designed to do what they bring in." EMA

"The weakness ... is that it is not seen ... or appreciated by the customers and even by QCA ... the other weakness is ... the, if you like, benign acceptance by schools of English is different from everything else ... the underlying thing is ... that, ‘because you allow us to [request R3 reviews], we believe there must be something wrong with [KS3 English marking in general] and because our profile is affected, we're going to find something wrong with it’. Whereas, [if forced to use R2s, they would] really have to interrogate individual pupils’ responses and say ‘that mark scheme wasn’t applied fairly in this particular case for this reason’ as they have to do for science and maths. Right? Then it doesn’t put English marking at a disadvantage; it doesn’t ... threaten it in terms of public image.” EMA

"Every decision you make there's paperwork around it and that does slow you down. If we can look at ways of saying ‘okay, what of this can we cut out?’." EMA

“a lack of consistency across the EMAs.” LCM

“ Weaknesses, as we've already said, are three different EMAs, three different sets of practice and, although consistency has improved over the years, there are still some inconsistencies in the ways that the reviews are carried out and that leads, I think, to slight differences in the quality of the process.” LCM

6.3.9.6 Most immediate scope for improvement

Finally, the most immediate opportunities for improvement were felt to be as follows.

“I think the way ahead is certainly looking at the way things are marked, looking at electronic transfer of data, looking at scanning reviews in to get them done quicker ... but the important thing is you maintain the credibility of the system you've got when you're doing it and that's the key to it. And again the time window.” DFEE

“even within the current cycle we need far more clarity. Having that clarity, we would want a lot more consistency across the nation because that's part of the underpinning principles of this new organisation, this new approach. And with that consistency, though, we do need that flexibility in terms of the ability to be responsive to situations. We don't just want to be agents of delivery, we also want to be agents of change ... I think improvements, in terms of reviews, are to do with communication, both with customers but also within the people who are working

nfer
across the three regions and with QCA; but you can only improve communication when there is sufficient clarity in the messages and the key points and the principles and so on.” EMA

“the establishment of the requirements on schools when making the re review, the absolute clarity on the EMA’s part about how they are going to do it.” EMA

“I think I would like to understand and address the interface with NDC, whether it actually is necessary to do R1s ... obviously NDC couldn’t pick up if there was ... additional mark on the scripts or wrong transfer, but I assume NDC would pick up that [there] were wrong additions on a mark sheet, wrong ... levels-to-marks.” EMA

“the ... review should be the full review of the whole work.” LCM

“The major scope for improvement is — building on that personnel issue — is to pull the thing into one centralised process ... a team of reviewers, senior personnel, led and monitored by the Lead Chief Marker. The timing for that is obviously going to be quite problematic ... and, as I say, it might have a cost implication; but I do think it will improve the consistency and the quality of what goes on.” LCM

“it’s got to be done in the same way whichever part of the country your school’s in” LCM

“I would look to next year just being a year of consolidation ... I think the EMAs have made a lot of changes this year, last year, to get the system to the efficiency which we clearly achieved this year.” NDCA
Section 7 An evaluation of external marking review services during 2000

7.1 Introduction

Studies 1 to 5 addressed the aims of the evaluation project from a range of different perspectives. Study 1 considered review procedures as they were intended to function during 2000 and Study 5 explored what key professionals within the system thought of these procedures. Study 3 considered the views of mainstream schools that had participated in the national curriculum testing process during 2000 (but that had not necessarily requested reviews), while Study 4 focused exclusively upon schools that had requested R2 or R3 reviews. Study 2 addressed the overall evaluative aim through an exploration of data relating to the review services from 1997 to 2000.

As explained in Study 1, the five studies are to be seen as a whole, rather than as free-standing entities. As such, the purpose of the present section is to integrate the various themes that have already emerged; in effect, to conduct the evaluation.

Before embarking upon the full evaluation, certain broad conclusions will be presented. These concern the overarching question of whether the review system operated effectively, consistently and fairly during 2000. They focus upon the roles of the External Marking Agencies and of participating schools, respectively.

7.1.1 Did each of the EMAs run a satisfactory operation?

Study 1 explored, in some depth, the various review procedures as they were intended to be operated by the EMAs during 2000. Although it was not possible to conduct a detailed analysis of whether these procedures were always effectively put into practice, limited evidence of implementation was also considered. From Study 1, the impression of the intended and actual functioning of the system was positive. Each of the three EMAs demonstrated a clear commitment to dealing with review requests in a manner that was responsive to the needs of individual schools and consistent with the requirements of the QCA. The impression from Study 1 was that review procedures were generally effective and fair and operated with sufficient consistency between the three English EMAs.

Furthermore, when a range of professionals involved with the system were interviewed during Study 5, they too indicated a general level of satisfaction. There was a feeling amongst these participants that — while there had been problems in the past — the functioning of the external marking and review system had improved considerably in recent years, with strong management from QCA and with good will from all parties.
involved. Above all, though, these participants stressed that the quality, the experience and the commitment of senior markers was the root of the success of the review system.

In terms of the quality of customer service provided by the EMAs, Study 4 demonstrated that most schools were satisfied. Generally speaking, of those schools that had made contact with EMA staff during the review process, most were happy with the level of service that they had received; the majority agreed that staff were polite, helpful and efficient and disagreed that they were unnecessarily bureaucratic.

Furthermore, from Study 2, it was clear that the majority of R1 and R2 reviews were turned around within specified time limits, although the targets were somewhat less well met with respect to R3 reviews. Typically, though, even when targets were not met, they were not exceeded by a large amount.

While the EMAs appeared to function with sufficient consistency, improvements could undoubtedly be made. As acknowledged by most of the Study 5 participants, there is a need to continue co-ordinating the work of the EMAs.

The NFER was charged with identifying any differences between EMAs that might lead to outcomes for schools which were materially different. Although, from Study 2, significant differences were observed between EMAs in the percentages of successful review outcomes, no one EMA emerged as consistently most lenient or most harsh. Moreover, there was no evidence that would allow the differences in outcomes between EMAs to be unambiguously attributed to differences in operational practice.

When, during Studies 3 and 4, impressions of the review system were analysed according to the EMA responsible for marking respondents’ scripts, only a small number of significant differences were observed. However, when differences did occur, they were almost always in favour of AQA-N. Moreover, they typically related to key stage 3 English at AQA-N. While there are problems in comparing responses between EMAs, it may well be that some of these significant differences can be attributed to the fact that AQA-N is the lead agency for English and that the Lead Chief Marker for KS3 English made a considerable contribution to the processing of requests at AQA-N.

7.1.2 Did schools use the system in a satisfactory manner?

The fairness of the review system is not simply dependent upon the extent to which valid procedures are operated consistently between EMAs. It is also imperative that the system is used effectively by schools. The issue is twofold. Firstly, are all pupils whose scripts are not marked to an acceptable standard submitted for review? Secondly, do schools refrain from requesting reviews for pupils whose work is marked to an acceptable
standard? In many ways, answers to these questions posed the largest threat to the validity of operation of the review system during 2000.

To begin with, it was clear that differences between schools were likely to have had an impact upon the likelihood of poorly marked scripts being submitted for review. From Study 3, it appeared that between 1% (KS3E) and 11% (KS3M) of schools did not check the marking of their pupils’ scripts at all. As such, there would be no way in which potential marking errors could be detected and pupils in these schools might be disadvantaged as a result.

There were also differences in the extent to which survey respondents were prepared to tolerate a small amount of marking error; moreover, the review-requesting schools of Study 4 were less likely than the sampled schools from Study 3 to tolerate such error. Again, then, this is a factor that might disadvantage some pupils and advantage others.

Likewise, across samples, there was a tendency for schools to request reviews repeatedly from year to year. Across the seven Study 4 samples, the percentage of the 2000 review-requesting schools that had also requested some kind of review during 1999 ranged from 1½ times to over 4 times the national percentage (i.e., across samples, the percentage of respondents whose schools had requested reviews during 1999 was at least 1½ times higher than the percentage of all schools that had requested reviews during 1999). This suggests either that some schools are requesting reviews more frequently than they ought to, or that some schools are requesting reviews less frequently than they ought to, or both. (It is assumed that the examiner-allocation programs employed by the EMAs rule out the possibility that an improbable percentage of schools really do receive consistently poor markers from year to year.)

One of the most significant, although perhaps least surprising, of findings was that schools tended not to request reviews when they believed that pupils had erroneously been awarded too high a level. This was despite the fact that a considerable percentage of respondents from Study 3 explicitly indicated that (at least one) of their pupils had been awarded too high a level as a result of marking error. In short, there seems to be an implicit or explicit tolerance of marking errors that work in the favour of pupils.

Finally, there was only a small amount of evidence that might support the proposal that a substantial number of schools were “playing the system” to maximise their performance table standing (for example, the fact that there were considerably lower levels of success as a result of reviews submitted by primary schools than by secondary schools).
Overall, then, there was evidence that the way in which schools made use of (or failed to make use of) the review system may have threatened its fairness for individual pupils. This is to underline the absolute necessity of ensuring that scripts are as accurately marked as possible before they are returned to schools.

7.1.3 The evaluation process

Having addressed the broadest of conclusions that emerged from Studies 1 to 5, the evaluation will continue by exploring the more focused issues in some depth. The aim of the following sub-sections is to integrate the various results from Studies 1 to 5, in an attempt to determine the extent to which the structures, procedures and practices employed during 2000 were optimal for the task of review.

The method adopted in the following sub-sections was explicitly to highlight potential weaknesses and scope for improvement. This was intended in the spirit of constructive criticism to support further developmental work. The fact that potential threats to fitness tend to be highlighted more often than current strengths should not be taken to imply that structures, procedures and practices during 2000 were unsatisfactory. As already explained, they were generally deemed to be satisfactory.

7.2 A framework for evaluation

For an evaluation to take place, criteria for judgement must first be established. As there was no pre-established set of criteria against which to judge the effectiveness of the review system, it was decided to construct one. To do so, the recently published *Arrangements for the Statutory Regulation of External Qualifications in England, Wales and Northern Ireland* were consulted. This document formulates and elaborates the standards by which the Qualifications and Curriculum Authority intends that UK qualifications should be judged. Four major themes were abstracted from this document:

1. interpretation and specification of ‘due process’, documentation of principles and of practices (i.e., how well is the system defined?);  

2. transparency, public understanding and user-friendliness (i.e., to what extent is the system accessible?);

---

1 The notion of ‘due process’ is borrowed from the legal context where it indicates the established rules and principles for the administration of justice. In the assessment context, while there is generally room for debate concerning precisely which methodological approach will best ensure the administration of a fair examination, there are many shared rules and principles that have been established over time that can be, and should be, drawn upon.
3. openness, independence and trustworthiness (i.e., to what extent is the system credible?);

4. technical effectiveness of structures and procedures (i.e., how well does the system function?).

The following discussion is broken down according to these four themes.

7.3 Interpretation and specification of 'due process', documentation of principles and of practices

Before considering the extent to which principles and practices were sufficiently well documented during 2000, we must consider a more fundamental question: exactly what purpose, or purposes, was the national curriculum review system designed to fulfil?

7.3.1 The rationale of the review system

Without an unambiguous and explicit rationale, a system cannot truly be said to be defensible; moreover, its effectiveness cannot fully be established. Unfortunately, underlying aims and purposes are often overlooked, or taken for granted, as systems develop and evolve. Indeed, it was not clear the extent to which the review system was grounded in an unambiguous and explicit rationale.

Implicitly, though, it appeared to be guided by at least four main principles:

1. to ensure that all pupils receive the national curriculum test levels that they deserve, given the quality of work produced in their test scripts (the measurement ideal);

2. to safeguard public and professional confidence in the external marking system (the political ideal);

3. to educate teachers concerning the appropriate interpretation of mark schemes and, by extension, national curriculum assessment objectives (the educational ideal);

4. not to distress, estrange or alienate teachers, pupils or parents (the psychological ideal).

To identify the foundational principles of the review system is not to engage in a 'purely academic' exercise. The relative value attached to each principle should – in practice as well as in theory – be central to any decision that concerns the system. This is most evident when principles come into conflict. For example, if most value was attached to safeguarding public confidence in the external marking system then this might
recommend different courses of action than if most value was attached to ensuring that all pupils receive the test levels that they deserve.

Yet decisions are not simply affected by conflicting principles, but also by practical limitations. Any decision will be mediated by a number of powerful constraints. The following (which, admittedly, are not independent) should be highlighted:

1. the cost of a particular course of action (a financial constraint);
2. the length of time required for a particular course of action (a time constraint);
3. the resources required for a particular course of action, i.e., qualified personnel (a resource constraint).

These purposes and constraints have been made explicit to frame the general evaluative discussion of Section 7, but also to draw attention to a number of specific issues that give particular cause for concern. These can now be highlighted.

7.3.1.1 Should pupils’ levels be lowered following a re-mark?

A general principle underlying the GCSE/A/AS enquiry-upon-results service during 2000 (and in previous years) was that pupils’ levels would not be lowered. This is explicitly not the case for the national curriculum review service. From the perspective of the measurement ideal, this is entirely justified. It appears also to be used as a threat to schools to help ensure that the system does not become over-burdened. However, the psychological impact upon a child whose level is lowered can be significant and this can also have a negative impact upon the public and professional perception of the external marking system. Thus, the measurement ideal comes into conflict with the psychological and political ideals.

Lord Mackay of Clashfern, in his 1999 report for Edexcel, addressed a similar conflict of ideals:

"The Code [of Practice for the GCSE and GCE] provides that on a re-marking as a result of such an inquiry the grade will not generally be reduced. While it is easy to see that the possibility of a grade being reduced may be a discouragement to make enquiries I question whether it is fair to the other candidates if, on a re-mark in accordance with the standard marking system applied, a script is found to have been over graded, it should nevertheless be left at that over grade because the matter has only come to notice as a result of an enquiry about results." (Mackay, 1999, p.25)
It should be noted that, in practice, pupils' levels were only infrequently lowered following review, especially following R2 review. The implication was that schools hardly ever submit pupils for review in order that their levels be lowered.

From 2001 onwards, the Joint Council has decided that GCSE/A/AS grades will be lowered, following an enquiry, should clerical re-checks or re-marks recommend this course of action. This would bring GCSE/A/AS procedures into line with those for the national curriculum tests.

7.3.1.2 Should re-marking be extended?

At GCSE/A/AS, when re-marking identifies a sub-standard marker, that marker's entire allocation is likely to be recalled, regardless of whether the schools involved have requested it. This very rarely happens as a result of national curriculum review, even when very inconsistent marking is detected that might well affect other schools. This is not consistent with the measurement ideal. Nor is it necessarily consistent with the political ideal: while a school might have its attention drawn to errors that it would not otherwise have noticed, it might, as a consequence, also come to respect the system more for having spontaneously dealt with the problem.

At present, the main argument for not extending re-marking would appear to come from pragmatic constraints: firstly, scripts would have to be requested from schools after the end of the summer term (which, in practice, would mean during the autumn term); secondly, more marking resources would be required, i.e., more markers or more work per marker (and marking resources are already scarce). However, whether these constraints alone are sufficient to defend the effective proscription of marking extension is an issue that deserves reconsideration. Alternatively, it may be that technological innovation can help to overcome some of the practical constraints, for example, through the scanning and storage of all scripts prior to initial marking.

7.3.1.3 Should only selected questions be re-marked?

Even for R3 reviews, it is not necessarily the case that the entire work of each pupil will be re-marked. For R2 reviews, schools are expected to specify where they believe marking errors have occurred and review markers are instructed to re-mark these questions. It is, therefore, quite possible for errors in other questions or papers to go undetected. Even worse, markers might notice errors but feel obliged not to address them. As was commented during Study 5: "often on the same script – even on the same page – there is another marking error that is ignored".

192
Where a marker has marked inconsistently, we might expect as many errors to work against a pupil as for her. If a school only highlighted errors in a pupil’s favour — and these were the only questions re-marked — then she would end up ‘winning on the roundabout’ without ‘losing on the swing’. This clearly elevates the psychological ideal above the measurement ideal. It is not clear how this is justified, other than in terms of the resources required to re-mark scripts. Again, whether such constraints are sufficient to justify proscribing full re-marks is an issue that deserves reconsideration. The same would apply for R3 review where scripts are not marked in their entirety.

7.3.1.4 Additional threats to the review ideals

The previous three questions highlighted ways in which principles can come into conflict. There are many other potential threats, though. The most obvious problem is the threat to the measurement ideal caused by schools (in the main) requesting reviews only when pupils’ levels are likely to increase. The somewhat counter-intuitive consequence of this is that the national performance profile, post-review, will be less accurate than it was pre-review (assuming that marking errors were not overwhelmingly errors of harshness). That is, the post-review performance profile will be a positively biased estimate of the performance of the nation (as only the mark scheme application errors that worked against pupils would be rectified). This is likely to corrupt the monitoring function of the national curriculum tests and to have a negative impact on attempts to set valid performance targets. Only for the individual pupils who underwent review would levels be more accurate post-review than pre-review. Once again, this highlights the absolute necessity of ensuring that scripts are accurately marked before they are returned to schools.

The measurement ideal is not simply threatened by schools failing to request reviews for pupils whose levels would be lowered, but also by schools failing to request reviews for a variety of other reasons. Studies 3 and 4 helped to clarify many of these, for example:

1. a lack of time to check pupils’ scripts fully;

2. a lack of time to complete review requests;

3. a perception that the results were of little importance.

Study 4 noted that the impact of the first two of these reasons was magnified for schools whose scripts were the worst marked (and who, therefore, would be most in need of marking review services). Clearly, this threatens not only the measurement ideal, but also the psychological ideal and the political ideal. Schools should not feel that the system is constructed so as to deter them from requesting reviews — particularly not those schools
that are worst affected by marking error. Indeed, perhaps the review system ought to be underpinned by a fifth principle – the accessibility ideal – to ensure that no school, with a valid concern over the quality of marking of pupils' scripts, is dissuaded from requesting a review by undue procedural demands.

A further threat to the measurement ideal was identified within the data-checking process of the NDCA. When the NDCA uncovers an error that would lead to a level change, but that had not been identified by the relevant school through the review process, the NDCA makes no amendment. The apparent reasoning is, firstly, that NDCA data should match school data and, secondly, that the NDCA should not be pro-active in suggesting that schools appeal. (Incidentally, these errors are logged and sent back to EMAs to feed back into the marker evaluation process.) This means that pupil-level data which is subsequently sent to LEAs can actually include anomalies whereby, for example, a pupil's marks may imply a different level from that awarded. These levels are not even changed by the NDCA following a request from the LEA. This would clearly also pose a threat to the political ideal.

It seems that the reason for this problem lies with concerns over the timing of the data collection process. It would appear that, under the current DfEE publication schedule, there is insufficient time for the NDCA to notify schools of marking errors in order for them to appeal. As noted previously, is not clear that such time constraints are sufficient to defend tolerance of the problems that arise as an inevitable consequence. Moreover, it is not entirely clear why a formal review request should be required from a school anyway, that is, if the NDCA is confident that it has identified an explicit marking error that could easily be rectified.

7.3.1.5 Weighing values

We began by noting that the relative value attached to each of the principles underlying a system should, in practice as well as in theory, be central to any decision that concerns it. The implication of the foregoing discussion is that attention should be re-focused upon the underlying purposes of the review system. Moreover, consideration should be given to how the different purposes ought to be weighted. While this is no simple matter, it is recommended that the measurement ideal ought generally to be valued above the political or psychological ideal. The measurement ideal should be valued above the psychological ideal on utilitarian grounds; it should be valued above the political ideal on the assumption that it is precisely the measurement ideal that public and professionals alike require confidence in.
7.3.2 The formal documentation of procedures and practices

The document *Arrangements for the Statutory Regulation of External Qualifications in England, Wales and Northern Ireland* is, in itself, testimony to the importance that is now attached to the specification of procedures for external assessment within the National Qualifications Framework. It emphasises the need for formal documentation of assessment principles, processes and practices – an axiom manifest most clearly within the Code of Practice for the GCSE/A/AS.

7.3.2.1 Formal specification of the requirements upon EMAs

How well does the system for reviewing national curriculum test results meet this evaluative criterion? In many ways, it could do better. A particular problem seems to be the way in which review procedures have, for a number of years, tended to remain largely implicit – as though all participating agencies were assumed to be fully appraised of QCA requirements and of generally accepted models of best practice in the processing of review requests.

While certain important requirements were made explicit in Annex N of the Tender Specification for 1996-1999, similar details were not presented in subsequent specification documents. Which, if any, of the principles outlined in the 1996-1999 annex held for 2000? Which decisions recorded in emails of 2000 will endure for 2001?

This is not necessarily to suggest that the QCA ought to specify in precise detail *exactly* what procedures it expects EMAs to follow each year. However, it would seem reasonable for requirements to be formally recorded at some level of abstraction and systematically updated on a regular basis. This would help to ensure that the documentation of intended principles, processes and practices kept pace with changes that occurred in the field each year.

The relative dearth of formal documentation would seem to imply that all participating agencies unproblematically share a wealth of background knowledge, based upon substantial experience of the system. This would not only present problems for any new agency that did not have this experience, it would also present problems for any existing agency that underwent major personnel change. Indeed, this is precisely what Edexcel had to deal with during 2000. If EMA managers, particularly new managers, are to be expected to differentiate between procedural ‘requirements’ and ‘permissible variations’ there must exist formal documents in which these parameters are made explicit. This was often not the case for 2000, for example, in relation to the training, re-standardisation and monitoring of review markers. Because of this, it was not always clear the extent to
which ‘due process’ in the processing of review requests actually existed – did it exist despite not being formally recorded, or did it simply not exist at all?

For 2001, QCA intends that the *External Marking Product Descriptions Compendium* should constitute one of its principal reference documents. It is not clear, however, that this will be sufficient to convey all procedural requirements. If processes, as well as products, are to be proscribed then it would seem appropriate that they be formally recorded (whether in broad terms or more precisely). Perhaps a version of Annex N, from the 1996-1999 Tender Specification, should be formally updated each year?

### 7.3.2.2 Formal specification of the requirements upon EMA staff and markers

It would be false to claim that there was no explicit specification of due process and it is important to highlight, in particular, the generic documents produced by AQA-S in consultation with the other EMAs, the QCA and various other stake-holder representatives. These included the *Schools’ Guide*, the *Markers’ and Supervising Markers’ Handbooks* and, to a lesser extent, the *Instructions to Markers Conducting Re-marking*. These generally presented clear and explicit interpretations of due process and facilitated the consistent application of best practice.

Yet, there is still scope for improving practice, for example, in the documentation given to markers (which was not consistent between EMAs during 2000). The Joint Council for General Qualifications produces generic booklets concerning procedures for post-examination services, not only for schools, but also for examiners and examination board staff. It is recommended that the QCA consider this approach. (Of course, the production of generic booklets would not prevent each EMA from providing its own additional instructions for markers or staff.) This might help to overcome some of the procedural confusion that occurred during 2000, for example, the classification of KS2 English review requests. Moreover, if these documents were made available in the public domain, this would help to make the system more open, transparent and accountable.

Within each EMA, during 2000, efforts were made to specify due process in the administration of review requests. Desk instructions were important for ensuring that staff completed their work consistently, appropriately and to deadlines. Quality Assurance statements, where these existed, were particularly useful in explicating some of the more general principles underlying the review system.

Likewise, it was clear that efforts were made to record departures from intended practice; toward this end, complete records on the processing of each individual request were kept
by each EMA. These would prove essential for auditing purposes or when schools challenged the processing of a request.

7.3.2.3 Levels of documentation

There are strong arguments in favour of the formal documentation of principles and intended practices at a number of levels, for example:

1. technical specification (full documentation of principles and intended practices – for internal use and accountability);

2. Code of Practice (general description of structural and procedural standards – for external use and accountability);

3. generic published instructions (cross-EMA guidance for implementing the Code of Practice – targeted at specific groups, i.e., schools, EMA staff and markers);

4. specific published instructions (optional EMA-specific guidance for implementing the Code of Practice – targeted at specific groups, i.e., schools, EMA staff and markers).

7.4 Transparency, public understanding and user-friendliness

It is not sufficient that procedures are well defined and formally documented, it is also important that they are accessible to professionals within the system and comprehensible to interested observers.

7.4.1 Transparency and public understanding

The consideration of transparency provokes a number of questions in relation to how well review services were understood during 2000, for example: Was there information to be found? Was it accessible? Was it clear? Was it detailed? And was it comprehensive? These questions are relevant in relation to any stake-holder who may have wanted to seek information, for example: pupils, parents, teachers, LEA officers, markers, EMA staff, QCA staff, NDCA staff, DfEE representatives, etc.. They are of greatest significance, though, for those stake-holders with direct involvement.

Once again, explicit documentation features as an important factor. We have already considered documentation for staff working within the system and later sections will consider documentation for schools in greater depth. However, it was apparent that little in the way of information on review services was provided for pupils and parents, meaning that teachers would have had to have shouldered this responsibility. Perhaps information on review procedures should be made more widely available? One way of
doing this might be to exploit the omnipresence of the internet. While not yet truly accessible to all, it is a cheap and effective method of making important information available in a variety of formats for a variety of audiences. This is one way of moving towards a greater public understanding of the system, a goal to which the QCA is clearly committed. While the QCA and DiEE made limited reference to review procedures on their web-sites, this information was intended primarily for schools and not for other stake-holders.

It is important to realise that the provision of information that is accessible, clear, detailed and comprehensive is not sufficient to ensure public – or professional – understanding of the system. Even when provided with information, members of the public and professionals alike must be motivated to process it and must be prepared to reject false preconceptions. If more and better information on review services is to be provided, it must be provided effectively to ensure that it is appropriately received. Determining how to achieve this goal is an important task that will require careful consideration.

7.4.2 User-friendliness

Linked to transparency is the issue of user-friendliness: how easy is it to make use of review services?

7.4.2.1 The time and effort required to submit review requests

Study 3 explored the extent to which various aspects of the system discouraged schools from requesting reviews. It found that ‘the complexity of the R2 review procedure’ was given the highest ‘discouragement’ rating (5 – very significant) by between 9% (KS3M) and 25% (KS3E) of respondents across samples. Between 26% (KS2) and 43% (KS3E) of respondents, across samples, gave ‘the time taken to check scripts and prepare an R2 review case’ the highest ‘discouragement’ rating. A large percentage of schools may, therefore, have been put off requesting reviews by procedural demands that they perceived to be too great.

The KS3 English respondents appeared to be the group most discouraged by procedural requirements. Studies of marking reliability for the O-level and GCSE examinations have consistently shown that certain subjects are more prone to problems than others. These tend to be subjects where there is more scope for interpretation and evaluative judgement and where the mark scheme is less clearly broken down and defined (Newton, 1996; Murphy, 1982 and 1978). English, especially, tends to be among the least reliably marked of subjects. As such, in the context of the national curriculum review system, it seems
likely that those worst affected by marking error were those most discouraged from requesting reviews.

Study 4 explored the user-friendliness of review procedures in more depth. Generally speaking, it found that review-requesting respondents were satisfied with the procedures that they had been required to follow. For example, it found that respondents tended not to have trouble locating a copy of the *Schools’ Guide* and that they tended not to agree that too much paperwork was involved in the review request. Yet satisfaction was not universal. Around 10% of key stage 3 respondents did indicate having had problems locating a copy of the *Schools’ Guide*. (Note that examinations officers were rarely involved in the process, which may help to explain this fact.) Furthermore, one-half of KS3 English R3 respondents and two-thirds of KS3 English R2 respondents agreed that there was too much paperwork. As before, then, those worst affected by marking error were those least satisfied by the procedural requirements for getting their errors rectified.

Across each of the seven Study 4 samples, most concern was raised with the statement ‘we had insufficient time to complete our review request(s)’ – with the exception of KS3 maths, between one-third and two-thirds of respondents, across samples, agreed with this statement; indeed, 37% and 23% of KS3 English respondents (R2 and R3, respectively) ticked the ‘Strongly agree’ statement. If the review system is to be fair, then it must be equally accessible to all. The fact that the system was deemed to be unreasonably demanding (in terms of time and effort) for precisely those schools worst affected by marking error is reason to believe that it was failing in a crucial respect. It is recommended that consideration be given to reducing the amount of time and effort required to prepare review requests.

Notably, the GCSE/A/AS enquiry-upon-results procedure places a smaller burden upon schools. More time is available to them and less form-filling is required. In particular, there is no requirement upon schools to explain exactly how they believe the marking of scripts to be in error (schools have traditionally not had access to scripts). While adopting this approach for national curriculum reviews would mean that all scripts would have to be re-marked in full, it would also mean that less time and effort would have to be expended by schools. Considering the amount of time and effort that would still be required of schools (to check the marking of pupils’ scripts) any attempt to reduce additional burdens would presumably be welcomed.

From the interviews of Study 5, there was evidence of a concern that reducing the burden upon schools might lead to a rise in the number of inappropriate review requests; for example, if schools no longer had to explain why they believed the marking of a script to be in error, they might be less inclined to check scripts appropriately and to rely more on
‘gut reaction’. This is a possibility. Reducing the burden upon schools might be expected to lead both to more appropriate review requests and to more inappropriate review requests. A central question, therefore, might be whether the cost of more inappropriate review requests would outweigh the benefit of more appropriate review requests. A more important question, though, would be whether inappropriate reviews could be discouraged in other ways. The loss of a deposit, of course, is one such mechanism.

7.4.2.2 The deposit

It was not clear, from talking to those involved in the system, the extent to which the deposit was intended explicitly to function as a disincentive to requesting reviews. From Study 3 and Study 4, though, it appeared to function as such for a considerable number of schools. Even at the present rate of £5 per pupil, between 19% (KS3S) and 23% (KS2) of respondents, across the Study 3 samples, gave this factor the highest ‘discouragement’ rating (5 – very significant).

If schools with genuine concerns over the quality of marking of their pupils’ scripts were genuinely being put off requesting reviews by the £5 deposit, then it would be appropriate to question whether there should be a deposit at all. On balance, though, the need to provide some deterrent to the inappropriate request of reviews would support the use of a deposit.

Whether the mechanism should be referred to as a ‘deposit’ is another question, though. One argument for instead calling it a ‘fee’ relates to impression management: the term ‘deposit’ gives the impression that it will be returned, which is almost an implicit expectation of error on behalf of the external marking system. On the other hand, this might be considered an appropriately empowering default expectation. Another argument for calling it a fee relates to the proposal that it should be treated as a payment for services rendered and, as such, set at a rate that corresponded to the real costs involved. Although increasing the cost of the deposit is not necessarily recommended, if this were to happen then relating the revised figure to the real costs involved would appear to be a defensible approach to take. It would also be worth considering whether a deposit (or a fee) should also be levied for R1 reviews (given the considerable number that failed to result in a change of level during 2000). Finally, it is important to stress that there is a fine line to be drawn between an appropriate deterrent function and an inappropriate one and any decision to raise the deposit rate should be guided by this consideration.

Crucially, referring to the deposit as a ‘fee’ in no way implies that it should not be refunded if marking errors are acknowledged – it should be. However, there are strong reasons to believe that the present refund arrangements are inappropriate. It is
recommended that, following the GCSE/A/AS precedent, deposits should be refunded only for those pupils whose levels change (i.e., on a pro-rata basis relating to success of outcome for each pupil). There is no obvious justification for not adopting this practice. Moreover, the present refund arrangements provide an incentive to submit inappropriate or marginal review requests in addition to appropriate ones.

Finally, it would be worth considering the use of a post-review invoice system rather than a pre-review deposit system. This would avoid delays caused by schools forgetting to enclose cheques and would mean that schools would have less to do to prepare their review requests. A variety of invoice models are employed at GCSE/A/AS and would seem to function effectively. However, there would be a need to consider the extent to which schools would be prepared to pay post-review invoices and what sanctions could be imposed if they did not. It would be important to ensure, in advance, that the system would be workable in the national curriculum context.

7.5 Openness, independence and trustworthiness

It is not sufficient that procedures are well defined, formally documented and accessible, they must also be credible and inspire confidence. The system must be – and must be seen to be – open, independent and trustworthy.

7.5.1 Openness

Explicit documentation is not simply important for facilitating the consistent operation of a system, it is crucial for presenting the system as an open one. It is not easy to determine whether a system is meeting its obligations if these are not clearly and explicitly recorded in advance. The openness of the review system would, therefore, be enhanced through fuller documentation. The more widely this documentation was made available, the better.

The NFER’s experience of researching the present project is, in itself, important evidence that the review system is open to scrutiny. However, the caveat should be added that this research was commissioned by the QCA primarily as an (independent) internal evaluation for developmental purposes rather than as an (independent) external evaluation for accountability purposes.

7.5.2 Independence

Issues of independence arose on a number of occasions during Studies 1 to 5. They can be discussed in relation to the EMAs and markers, respectively.
7.5.2.1 The independence of the EMAs

The very idea of calling upon an agency to validate an alleged error for which it might have to admit full responsibility may not promote public or professional confidence. Are rejected requests simply attempted cover-ups? It is not hard to see how an aggrieved school might see it that way. This problem is compounded by the fact that there is no independent body with which appeals may be lodged.

Three alternative models were suggested.

7.5.2.1.1 Single EMA with national curriculum markers

Under present arrangements each EMA deals with its own review requests. An alternative approach might be for the Lead Agency for each subject to deal with all review requests, regardless of which EMA was responsible for the initial marking of scripts. At least for two of the three EMAs, then, this might give a better impression of independence (although, this is debatable for the two AQA sites). Moreover, it would facilitate far more effective communication between EMAs, LCMs and markers and would help to enhance consistency of review decisions. Of course, it would mean considerably more review work for the Lead EMA for English and considerably less for the Lead EMA for maths.

7.5.2.1.2 Wholly independent agency with national curriculum markers

An alternative would be for an entirely independent agency to take on the task of processing review requests, for example, the DfEE. It would still appoint the same senior markers though. The principal advantage is that it would give a clear impression of being independent. However, using entirely new managerial staff would risk undermining consistency with the rest of the system. The financial constraint of such a system would also need to be considered.

7.5.2.1.3 Wholly independent agency with wholly independent markers

The most extreme possibility would be for an entirely independent agency to take on the task of processing review requests using an entirely independent team of markers. This would give the clearest impression of independence (although marker commitments would probably not be transparent under the second model anyway). However, it would run the risk of introducing unnecessary inconsistency in marking. It is hard to see who could be trained to be as effective re-markers as the most senior markers from the initial administration.
7.5.2.1.4 Which model is most appropriate?

In many ways, moving to one of these alternative models would be constructive. Not only would the appearance of independence be stronger, but the management structures would be tighter and the communication channels would be clearer. One concern that was raised came from Chief Markers who felt that the present system allows for a final check on the quality of initial marking within an EMA. Whether this would be an insurmountable problem under each of the alternative models is not clear.

If required to choose between the three models, the NFER would recommend the first. This would maximise the appearance of independence while minimising risks to the technical rigour of the system.

7.5.2.2 The independence of the Lead Chief Marker

In previous years, Lead Chief Markers had no formal national role in the co-ordination of review re-marking; they were appointed as Chief Markers within EMAs which is where their primary responsibilities remained. For 2000, LCMs were no longer to act as Chief Markers within EMAs but were to have a more independent role.

Unfortunately, the revised role of the LCM in relation to the processing of reviews was less than clear during 2000. Indeed, the roles adopted appeared to owe more to tradition and personal circumstance than to job specification or logic. One problem that arose was a perceived conflict between the principle of independence and the reality of being appointed by one of the EMAs. It was unclear whether the role of the LCM gave her authority over the Lead EMA that appointed her, or over either of the other agencies. Likewise, it was unclear whether the other EMAs could expect the same level of involvement from the LCM as the Lead EMA. These ambiguities need to be resolved for future years. Indeed, it is worth considering whether the LCM should be a QCA appointment, to provide the necessary level of independence.

There are more general issues that need to be resolved in relation to the role of the LCM. Firstly, review responsibilities need to be reconsidered. This will involve considering whether current expectations for national monitoring are realistic for what is typically a part-time post. Indeed, general levels of involvement with the review process will need to be debated as these differed markedly between subjects and key stages during 2000.

Issues of communication need also to be addressed as concern was evident that channels between QCA, EMAs, LCMs and CMs were not functioning as effectively as possible. Particularly if the LCM is to be genuinely independent of each EMA, it is essential that she be kept reliably informed of all relevant decisions and progress. During 2000, there
was a feeling that communication channels did not always function as effectively as they might between QCA and the LCMs, CMs and DCMs – the problem being the EMA ‘filter’. This issue will need to be addressed for future years.

7.5.3 Trustworthiness

Even when a system is open and faultlessly independent, it may still fail to inspire public and professional confidence. This may be justified, when procedures are not appropriate in principle or when they fail in practice; but it may also be unjustified, when invalid criticisms capture the limelight or when impressions are coloured by dissatisfaction related to other systems. It is likely that each of these factors go some way to explaining concerns that have been raised with the review system in recent years. The following subsections explore levels of trust vested in the review system by members of the teaching profession.

7.5.3.1 The public face of professional confidence

Issues of public and professional confidence in the review system are at their most stark when represented in the media. The following quotations were extracted from letters to the Times Educational Supplement during 1999.

“A ‘rogue’ marker was blamed for the errors, but I can’t help wondering how many other schools have received depressed (and depressing!) results and not checked the scripts against the mark schemes to see if there are any mistakes or meanness in the marking.” (T Purcell, TES, 29/01/99)

“In response to the QCA’s assertion that so few scripts are re-marked, which must show how accurate they are, I know a number of schools who also felt that their tests were very badly marked, but who did not appeal because of the very tedious process and the feeling that these results do not matter to anyone.” (P Dearden, TES, 26/02/99)

“We also sent back many of our Year 9 English papers. The results we have now received bear no relation to the originals.” (C Gaskin, TES, 08/10/99)

While these comments came from 1999, they reflect ongoing problems stemming back to the introduction of the tests and, in particular, to the introduction of external marking. Similar concerns were raised during 2000. During September, the Times Educational Supplement ran an article entitled “Marking blamed for thousands of reviews”. It noted:
"Queniborough primary, in Leicestershire, returned half of its English papers and all were upgraded, increasing the number of children reaching the expected level from 33 to 56 per cent." (TES, 29/9/00)

In fact, according to Chris Davis – NPhA Media Correspondent and Head of Queniborough Primary – 100% of the pupils' marks were increased and 78% of the levels (comment on NPhA website ‘Staffroom’ dated 8th September 2000). Whatever, the figures do not make for encouraging reading. Importantly, though, they do make for good reading, at least from the perspective of the press. As long as the review system is in existence it will be possible for journalists to find stories such as this. Effective media management is, therefore, crucial.

In order to rebut claims of poor initial marking – in the face of high-profile review 'successes' – the QCA has typically adopted the strategy of explaining how small the percentage of pupils involved in reviews is. This is, no doubt, appropriate. When faced with criticisms of the extent of marking errors in GCSE and A-level scripts, Edexcel used a similar approach, in addition emphasising that clerical re-checkers were closely monitored:

"An internal audit has shown that the margin of error is incredibly small because there are so many checks and balances." (Quoted in The Sunday Times, 10/09/00)

Media management is a very tricky business, though, and can run the risk of backfiring, thereby leaving the last word to others. This happened to Nick Tate (then Head of QCA) during 1999, following the publication of an article in the TES entitled "Tests can only get better", in which he defended public tests and examinations, describing the national curriculum test system as "ambitious and sophisticated". Two weeks later, a letter from Barrie Day entitled "Would you buy a test from this man?" was published. It contained the following evaluation:

"This smug, self-congratulatory piece about the curriculum testing regime being a 'high-quality product' in which the 'level-setting procedures are some of the most sophisticated in the world' was more the musings of a fiddling Nero than an informed assessment of the true reality for the humble consumers of the 'product'."

"Two years ago our English test papers were ridiculously over-marked – we gained 81 per cent level 5 and above. This set up totally false expectations of future achievement in the minds of pupils and parents.

nfer 205
But then 1999 was the year of the hatchet job. We gained only 47 per cent level 5 and above. This was unbelievable with a similar cohort of pupils as previous years and improved teaching. So after 10 hours of scrutinising our pupils’ test papers, we challenged this obvious travesty. The results of the re-mark? Of the sample of 47 pupils remarked, 27 were raised a level – a 10 per cent change! One girl was upgraded by 19 marks – almost two levels! And what of the other 63 pupils who were not re-marked? We’re pursuing it. Is this what passes for sophistication Dr Tate?” (B Day, TES, 24/09/99)

This illustrates, once again, how a single case of poor quality marking – backed up by an effective review – can have a national impact. However, it also illustrates the risks involved when mounting a high-profile defence of the testing system, particularly from within that system. This again highlights the importance of independent external evaluation.

7.5.3.2 The private face of professional confidence

Study 2, Study 3 and Study 4 explored more systematically how schools used and perceived the review system.

7.5.3.2.1 Perceptions of the quality of initial marking

As discussed during Study 2, although only a small percentage of pupils are involved in review requests each year, a considerably larger percentage of schools are. Particularly at key stage 3, where one-third of schools submitted some kind of review request for English and one-fifth for science, the profile of the review system will inevitably be high. Moreover, with large proportions of all review requests resulting in level changes, the review system will present large numbers of schools with formal confirmation of their concerns regarding the quality of the initial marking of scripts.

In fact, there is some injustice in the impression given by the refund of a deposit following a successful review. From the quotations presented above, it is clearly taken by schools as an implicit – if not explicit – acknowledgement by the EMA that the standard of initial marking was not up to scratch. Yet an increase in the level of even a few pupils, from a school with a sizeable cohort, is by no means a valid indictor of unacceptable marking error. Even when marking is carried out to an extremely high standard, occasional errors of the size of only one mark can lead to post-review upgrades when those errors occur for pupils just below level thresholds. The possibility that formally acknowledged marking errors may be perceived as admissions of poor quality marking –
even when the overall quality of marking was good – is a further impetus to ensure that initial marking is as accurate as possible before scripts go back to schools.

Furthermore, it raises the question of whether there is more to be done to educate schools in the notion of ‘acceptable tolerance’ in the marking of scripts. Indeed, there may be an argument for publishing more than just bare data concerning the numbers (and percentages) of reviews that result in level changes for pupils. A more valid index would be to publish the proportion of re-marks that led to mark changes within or beyond pre-determined thresholds. This would give a far better indication of the extent to which initial marking was effective. Of course, this would not be an absolute index of effectiveness, as schools do not submit poorly marked scripts for review where there is no expectation of a level change (and some schools do not submit poorly marked scripts for review even when level changes would occur).

Not surprisingly, the most severe reservations over the quality of initial marking were noted in relation to English. Yet it is important to note that KS3 English fared particularly badly – perhaps more so than might be expected. From Study 3, only 55% of KS3 English respondents believed that the quality of initial marking was satisfactory (or better), which meant that 45% believed the quality to be poor (or worse) – and these results came from a sample of all participating schools, not simply those that had requested reviews.

Interestingly, from Study 3, there was a greater perception of marking error against pupils than in their favour (across all samples). Unless marking in all samples was genuinely more harsh than lenient, this suggests a subtle problem – it suggests that schools generally perceive the marking system to be loaded against pupils.

7.5.3.2.2 Perceptions of the quality of review re-marking

When the Study 3 samples of regular, participating schools were asked their opinions of whether the review system was generally fair and sufficiently independent and impartial the response was cautiously positive. While the majority were in agreement, around one-third (across samples) ticked the ‘Undecided’ box.

Similarly, when asked whether they believed that most scripts ended up with the correct level following R1 or R2 review, these schools generally did not disagree. However, for R1, between one-fifth (KS3S) and two-thirds (KS3E) were ‘Undecided’; for R2, between one-third (KS3M) and one-half (KS2E) were ‘Undecided’. This is not a particularly reassuring outcome. It seems to suggest either that the respondents did not feel they had enough information to decide, or that the information that they were in possession of was
indeterminate. Either way, there is a clear need for some kind of public relations work in this area. In the same way that marking reliability studies are essential for upholding confidence in the rigour of initial marking, there is a strong case for the same kind of study applied to review marking.

When the Study 4 samples of review-requesting schools were asked whether the review system was fair, independent, transparent and accountable they were generally in agreement. However, there were dissenting voices and these should not be ignored. They came particularly from KS3 English.

When these schools were asked whether the review system generally ensures that children receive the correct national curriculum levels for their work, at least 68% of respondents from the maths and science samples agreed. In contrast, only 59% of the KS2 English respondents agreed and no more than 43% of the KS3 English respondents agreed. Finally, while fewer than 9% of maths, science and KS2 English respondents agreed that they had lost faith in the review system, as many as 21% and 35% of KS3 English R2 and R3 respondents agreed.

Clearly there is a worrying pattern of disenchantment that affects KS3 English in particular and KS2 English to a lesser extent. This is apparent both with respect to the quality of initial marking and with respect to the quality of review re-marking. It is recommended that action be taken to address this lack of confidence.

More generally it was also somewhat worrying to note the way in which teachers tended to show no ownership of the marking and review system, despite the fact that so many of the markers were teachers. There was evidence of an ‘us’ (teachers) versus ‘them’ (the EMA/QCA) culture. This was particularly apparent from comments written on questionnaires, but also from telephone conversations with teachers. While this may be an understandable consequence of the way in which testing has been imposed on schools, it raises the question of what can be done to improve the situation.

7.6 Technical effectiveness of structures and procedures

Ultimately, the validity of decisions arising from the review system rests with the technical effectiveness of structures and procedures designed to ensure that pupils receive the correct levels for their work. An evaluation of these structures and procedures involves asking questions such as:

- do communication channels function effectively?
- do procedures maximise the trade-off between validity, reliability and manageability?
• are consistency and flexibility of response adequately balanced?

• do all ‘customers’ receive an equivalent level of service?

• are there effective mechanisms for quality assurance and quality control?

• do personnel possess a sufficiently high level of expertise?

• are personnel given sufficient instruction, guidance and support?

• is the system streamlined and cost-effective?

The following sub-sections will consider these kinds of questions in relation to the major stages of the review cycle. The discussion will be broken down according to the generic headings used in Studies 1 and 5.

7.6.1 Procedures for informing schools of their right to request a review

The principal method for informing schools of their right to request a review, for explaining review processes and for instructing schools on how to complete review requests was the Schools’ Guide. It was a clear improvement on the comparable document from previous years -- the Instructions to Schools -- and appeared to be well structured, well formatted and well illustrated.

Study 4 explored the extent to which schools (that had requested R2 or R3 reviews during 2000) were satisfied with this booklet. The majority indicated that instructions in the Schools’ Guide were clear (no more than 8% of respondents, across samples, responded with ‘Unclear’ or ‘Very unclear’). Furthermore, the majority of respondents, across all samples, also indicated that they understood how their EMA would undertake the review (although this impression was least positive for KS3 English). There was general disagreement that the Review Services section was hard to follow (although over one-quarter of KS3 English R3 respondents agreed that it was).

Generally speaking, then, schools were satisfied with the Schools’ Guide, although there was indication that improvements could still be made.

7.6.1.1 Instructions to schools on completing review requests

There are a number of specific issues that might be addressed in relation to clarifying the instructions given to schools. First, there was no advice on how schools should check pupils’ scripts. As this is one of the most important pre-review tasks, it may be worth
addressing the issue explicitly. This is particularly salient bearing in mind the range of checking practices that were observed from Study 3 and Study 4. Such instructions might, for example, suggest that schools: begin by comparing test levels with teacher assessment levels; continue by highlighting pupils for whom there is an unexpected discrepancy; then check the relevant scripts for clerical errors or mark scheme application errors.

Second, if schools are to be required to provide an explanation of 'how it is considered that the original marker has not applied the mark schemes correctly', then it is suggested that examples of acceptable and unacceptable justifications be included in the Schools' Guide.

Third, Study 4 showed that schools adopted different practices when submitting scripts for R3 review. It is recommended that schools be required to return all scripts from all pupils within their cohort.

7.6.1.2 Explanations for schools on how review requests will be dealt with

It is recommended that the Review Services section of the Schools' Guide be divided into two distinct sections. The second would present instructions to schools on preparing their review request(s), as described above.

The first would explain the system in considerably more depth than was attempted for 2000. It would describe the logic of the review process and would explain exactly how EMAs undertake reviews. It would emphasise the degree of rigour involved (including training and monitoring) and the fact that only the most experienced senior markers undertake re-marking. It could be used to 'sell' the review service as a system for righting potential wrongs.

7.6.1.3 A separate booklet

One key decision that should be considered is whether review services information should remain exclusively within the Schools' Guide. Alternatively, as for GCSE/A/AS, a separate 'post-testing services' booklet could be produced and this could be included with the return of scripts. This would help to ensure that appropriate members of staff had access to appropriate information at the appropriate time.
7.6.2 Procedures for the administration of review requests

The administration of review requests during 2000 was predominantly paper-based, with key data transferred to computerised databases within each EMA. Review requests are inevitably paper-based, as they relate to pupils’ scripts, and the tracking systems employed by each EMA appeared to function effectively. The computerised processing of data was somewhat more problematic, though.

7.6.2.1 Scope for enhancing consistency

While there were many similarities between the methods adopted by the EMAs, there still existed scope for enhancing the consistency of administrative procedures. As schools have no option for choosing the EMA to which they are allocated it is only fair that each EMA should treat review requests in a similar manner.

As explained earlier, this could be facilitated through more explicit documentation of intended practices. This might included clarification of the meaning of procedures; for example, how thorough a clerical check should be, or whether the requirement that all scripts received should be clerically checked should extend to scripts in R3 samples that are not re-marked.

There is also scope for enhancing the consistency of resources used, or documents prepared, by the EMAs. For example, the possibility of generic tracking forms might be considered, or the wider adoption of standard acknowledgement letters.

7.6.2.2 Structure and flexibility

It became apparent from Study 5 that there is work to be done in clarifying the approach taken to review requests that do not precisely conform to the guidance laid down in the Schools' Guide. The ‘zero tolerance’ approach, recommended by the QCA during 2000, was not followed to the letter. Indeed, there were good reasons for not doing so.

The ‘zero tolerance’ approach was recommended because of a perception that too many schools were abusing the system – by intentionally not following recommended procedures – with knock-on effects for the effectiveness with which all reviews could be processed. As was evident even from 2000, a few problem cases can cause undue obstruction and result in significant delay.

On the other hand, there will inevitably be instances where schools fail to follow procedural guidance for reasons that are justifiable. In these cases, refusing to complete a review request will not simply flout the measurement ideal (of ensuring that pupils
receive the levels they deserve) it will also flout the psychological ideal (of not alienating pupils or teachers).

There is a need for flexibility within structure. Moreover, if at all possible, slack should be built into the system to accommodate this. (With respect to the national curriculum review system, the main obstacle would appear to be very tight reporting deadlines.) However, there is also a need to set parameters which delimit the amount of flexibility that is permissible and this is the level at which debate is recommended for future years. It is recommended that this debate should be conducted amongst a range of stake-holders, including QCA, DfEE, EMAs, markers and teachers.

Finally, it is stressed that, where flexibility within structure does lead to departure from intended practice, these departures should be fully and formally recorded.

7.6.2.3 Review databases

Reconciling data from the three EMAs caused the QCA some problem during 2000. Gathering key data for Study 2 also proved problematic for the NFER. It would appear that work needs to be done to improve the ways in which data are recorded, stored and retrieved by the EMAs.

To some extent, this will require clear decision-making on the part of QCA management. Explicit and detailed guidance needs to be given on the way in which KS2 English and KS3 English R3 review requests are recorded. When data are published in future years, it will be important to explain the extent to which they can be considered comparable with historical data.

More general discussion also needs to take place concerning the data requirements of future years. In particular, it will be important to agree a policy on recording data relating to review commencement and completion dates.

7.6.3 Procedures for selecting, training and standardising review markers

Of the various stages of the review system during 2000, the stage of selecting, training and standardising review markers appeared to involve the most ambiguity. Few explicit guidelines for best practice had been laid down and differences in practice between subjects, key stages and EMAs were evident. Whether they resulted in materially different results for pupils was not clear and it is not suggested that they necessarily would have done. However, there is a need to address these inconsistencies. If differences are to remain, they must be fully and explicitly justified.
7.6.3.1 Markers

One model of best practice that had explicitly been specified in advance was that only grade A supervising markers should be relied upon for review re-marking. The EMAs were effective in meeting this criterion, despite problems such as the fact that many markers were teachers who were very busy at the end of term and who were in need of a holiday. Of course, full credit should be given to the markers, who demonstrated their thorough commitment through participation during this demanding period.

For future years, it may be worth reconsidering certain elements of the criteria for the selection of markers. As was pointed out during Study 5, all that is really necessary at this stage is precision of marking – supervisory abilities are generally not required. There are also a number of relatively minor issues that should be clarified, for example, whether the quality of the marker should be judged on the basis of the present year’s marking or on the basis of the previous year’s (the present year’s marking is unlikely to have been fully evaluated before reviews begin). Likewise, it should be considered whether a marker should be allowed to undertake review re-marking if a review against her own marking has been requested.

7.6.3.2 Training

To what extent do review markers need to be trained to undertake re-marking? Guidance was not explicitly given during 2000. All that markers received in the way of guidance was a brief instruction booklet which contained little in the way of detail (and which, for AQA-N and Edexcel, was generic across key stages and subjects).

Particularly where only the Chief Marker and the Deputy Chief Marker were involved in the re-marking of review requests (as was typical for maths at both key stages) it appeared to be taken for granted that no further training would be required.

Perhaps the most significant oversight in the system for 2000 was the lack of formal training to undertake those aspects of the process that differed from initial marking, i.e., explicit instruction on how review outcome reports should be written. If review outcome reports are to be written in future years, it is recommended that instructions for markers be included in relevant booklets and that formal or informal training sessions be considered.

A scenario-based training of markers, for KS3 English R3 sample decisions, was undertaken at AQA-N and Edexcel under the direction of the LCM (AQA-S decisions were not made by markers); this training appeared to function well.
7.6.3.3 Re-standardisation

Re-standardisation was, perhaps, the most ambiguous of all review issues evaluated during 2000. Neither principles nor practices were clearly defined. Although it had been recommended that re-standardisation be carried out, it was not actually clear what this should mean. Was it supposed simply to imply re-familiarisation with the mark scheme? Was it supposed also to require that a sample of 're-standardisation scripts' be marked – and the marking approved – before the re-marking of 'live scripts' could begin (as for the standardisation of initial marking)? Was it supposed to involve practice in, or the formal assessment of, report writing?

As above, where only the Chief Marker and the Deputy Chief Marker were involved in the re-marking of review requests (as was typically the case for maths at both key stages) it appeared to be taken for granted that no re-standardisation would be required. On the other hand (and particularly for English) where numerous markers were involved it was generally felt that re-standardisation was important. Moreover, it was felt that the LCM had an important part to play in co-ordinating this re-standardisation.

As was noted during Study 5, the decision as to whether re-standardisation is necessary should not simply be made on technical grounds. Also important is the political ideal of safeguarding public confidence in the external marking system. It is not only important that the system is fair and effective, but also that it is seen to be so. As such, there is a strong case for some kind of re-standardisation across the board.

It is recommended that re-standardisation procedures be developed for future years in all subjects at both key stages. However, it is not recommended that they should follow the formal, quality control model of initial marking, as all involved have already demonstrated very high marking standards. Instead, it is recommended that the process should require the re-marking of a sample of scripts (or questions from scripts) that have been selected by the LCM to represent a range of problematic issues that are likely to arise during the review period. There should be opportunity for the marking of these scripts to be informally discussed with the LCM prior to the commencement of re-marking. This would ideally be conducted at the beginning of a team re-marking meeting (see later subsections).

To avoid terminological confusion, it might be worth considering an alternative name for the re-standardisation process – a name that will help to distinguish it from the standardisation of initial marking.
7.6.4 Procedures for undertaking review marking

While procedures for processing review requests during 2000 were generally satisfactory, a number of issues deserved further attention. Particularly problematic were procedures for the processing of R3 review requests. These will be discussed below, with further comments on other procedural matters that arose.

7.6.4.1 Review turn-around times

EMAs were required to complete all reviews by the 22\textsuperscript{nd} of September 2000. Study 5 participants generally felt that this was fair and manageable; Study 2 showed that this target was met in the majority of instances. Concerns were expressed that the process would suffer if more stringent deadlines were imposed. It was felt that this could result in rushed work and might threaten the quality of re-marking. Especially bearing in mind that even the present deadline is, effectively, too late for pupils who have already left primary school, it would seem inappropriate to recommend tightening review turn-around times.

7.6.4.2 Re-marking workload

Some concern was expressed that there was unnecessary bureaucracy in the task expected of markers; for example, there was a feeling that there were too many forms to complete. Indeed, it was even suggested that having to carry out more work to reject a review (i.e., complete an outcome report) might act as a disincentive to reject a school’s request. While no evidence was presented in support of this possibility, it would seem sensible to work towards minimising the administrative load upon markers wherever possible.

One suggestion from Study 5 was for markers to be provided with word-processed templates on which they could record their comments. This would also reduce the administrative burden upon EMAs, as they (generally) type up hand-written comments.

Another suggestion might be to follow the GCSE/A/AS precedent and not give detailed explanations of why review requests were rejected. This point will be returned to.

7.6.4.3 The use of comment banks

The use of standard-response comment banks for science proved to have been an efficient labour-saving device during 2000. Moreover, there was no evidence that schools were less happy with outcome reports based on this approach than with outcome reports in other subjects.

While this model should be explored further, there were concerns that it might not be equally applicable in either maths or English. However, at AQA-N, an informal cutting-
and-pasting approach to report processing helped to speed things along. Whether this could be developed more formally might be worth exploring.

7.6.4.4 Re-marking monitoring (R1 and R2/3)

Like re-standardisation, the monitoring of re-marking was shrouded in an air of confusion and resulted in divergent practices between subjects, key stages and EMAs. Although there was a general feeling that monitoring ought to be taking place, there was no consensus over exactly how, or even what should be its purpose.

Monitoring can be understood from two main perspectives. Firstly, from a 'summative assessment' perspective, sampling the clerical checking and re-marking process can be considered an element of audit. That is, checking a random sample of the work produced during review provides a quantitative index of the technical effectiveness of the system. Secondly, from a 'formative assessment' perspective, sampling the clerical checking and re-marking process can be understood as an important mechanism for ensuring the quality of the product before it leaves the EMA. That is, any errors missed by the system can be rectified and further guidance can be given to staff or markers that are not performing to standard.

The second perspective is more akin to the (re-)standardisation procedure discussed earlier. In terms of the quality of the product that is returned to schools, it is of more significance than the first perspective. However, even from this perspective, whether there is a need for formal monitoring in addition to the re-standardisation of marking may be questionable. Indeed more informal monitoring of R2 and R3 re-marking may be sufficient.

A model that was often highlighted during Study 5 as an example of good practice was the team meeting approach of AQA-S. For all subjects and both key stages, AQA-S arranged meetings to which all those involved in the re-marking of scripts were invited. (For maths, only the Chief and Deputy Chief Marker were present; for KS3 English, over 30 senior markers were present.) These meetings were convened to complete as many of the re-marks as possible and stretched across a number of days where necessary. The real significance of these meetings was the opportunity for informal, collaborative peer-monitoring. Markers were encouraged to consult with each other, or with the Chief Marker, Deputy Chief Marker or the Lead Chief Marker (where present). The samples of work monitored were, therefore, not randomly selected, but selected on the basis of those that were proving most problematic.
It is recommended that this team meeting approach be adopted across EMAs to facilitate an informal, problem-centred approach to monitoring. More formal monitoring of re-marking could also be introduced if deemed appropriate (for example, each marker's first five scripts might be double-marked, and the advice of the Chief Marker sought where necessary). Where possible, it would be appropriate for the Lead Chief Marker to attend each meeting. Finally, it is recommended that each meeting be attended by at least two senior markers. The re-marking of an entire subject by a single, un-monitored, marker (as for KS2 maths and KS2 science at Edexcel during 2000) should be discouraged - even where it is the Chief Marker. For the sake of the political ideal, it is advisable that there be at least some evidence of re-mark monitoring.

There was somewhat more evidence of monitoring with respect to the clerical checking process during 2000, at least for AQA-S and AQA-N. AQA-S double-checked each script and AQA-N sample-checked at the data input stage. As a more mundane task, potentially more susceptible to lapses of concentration, it is recommended that a more formal approach to monitoring be adopted for clerical checks. It is suggested that all clerical checks be undertaken twice. If this is deemed not to be cost-effective, it is recommended that a random sampling approach be adopted with, say, 10% of scripts checked twice.

7.6.4.4 Should R3 reviews be restricted to KS3 English?

During 2000, the R3 review was available only for key stage 3 English. There appeared to be little in the way of theoretical rationale for this inconsistency; instead, the reason appeared to be pragmatic, i.e., that more substantial marking problems had been experienced for key stage 3 English than for other subjects.

It was clearly of some concern to a number of schools that they were not allowed to request R3 reviews in subjects other than KS3 English. They were faced with the alternative of requesting an R2 review for each individual pupil whose work was deemed to have been poorly marked. Where genuine large-scale marking blunders had occurred this would have required a great deal of time and effort.

It does not seem plausible that there were a substantial number of seriously errant markers for English at key stage 3, but none for English at key stage 2. If even a few schools are in need of an R3 review in subjects other than KS3 English then it would seem fair that this opportunity be made available to them. Indeed, the Schools' Guide should recommend this course of action where teachers feel that more than a specified percentage of the cohort has been awarded an incorrect level. (A percentage approach seems more
appropriate than the somewhat arbitrary ‘30 pupils’ currently specified which, thereby, recommends different tolerance levels for schools of different sizes.)

7.6.4.5 How should the R3 sample be selected

During 2000, schools adopted different approaches to the submission of R3 review requests. Some sent in scripts for the entire cohort and specified the worst 15 as their sample. Some sent in only 15 scripts and the remaining scripts had to be requested. Some sent in a sample of 15 and a ‘group’ of other scripts that had been poorly marked. Some simply sent in the scripts that they believed to have been most poorly marked, but did not specify a sample. Clearly, steps should be taken to eliminate this inconsistency.

Under current practice, a school is supposed to submit a sample of 15 pupils ‘whose scripts typify the perceived marking errors’. Clearly, then, this is a biased sample. In fact, it is entirely possible that a school with a large cohort will be able to furnish 15 poorly marked scripts even though the vast majority of its scripts were adequately marked. As such, it is possible that full re-marking exercises are being conducted that would not have been recommended had a random sample of 15 scripts been re-marked first.

In fact, as suggested by the senior QCA manager for external marking, this possibility could be investigated empirically from the 2000 databases. For schools that have had a full re-mark of scripts it would be straightforward to select random samples of 15 and – considering differences between prime and re-mark (perhaps even to a lower tolerance than for initial marking) – estimate the proportion that would not have been re-marked had the decision been based on those samples.

It is recommended that schools be required to submit all scripts from their entire cohort when requesting an R3 review. The decision as to whether a full re-mark should be conducted would then be based on a random sample of pupils (again, perhaps to a lower tolerance than for initial marking). This would bring national curriculum practice into line with GCSE/A/AS. It would be far more defensible from a technical point of view. Importantly, though, schools should also be encouraged to select out any pupils that they would still want to submit for an R2 review if the R3 review was denied.

Under this model, it is recommended that the sample be treated exclusively as a sample; that is, if a decision was made not to conduct a full re-mark, then – even where the sample re-marking led to a change of level for one or more pupil – levels should be neither raised nor lowered. A pupil’s level should only be changed where the school has specified in advance that the pupil should be subject to an R2 review if the R3 failed. Although this would go somewhat against the measurement ideal, it is recommended on the basis that
the sampled pupils should not have an unfair advantage or disadvantage in comparison with their peers (this would also be supported by the psychological ideal, should a sampled pupil be re-marked at a lower level).

Finally, there is a strong case for double-marking the samples of 15 (or, at least, any samples that did not lead to the decision of a full re-mark). Not only would this serve the measurement ideal, it would also serve the political ideal when explaining to schools why their R3 request was not granted.

7.6.4.6 R3 borderlining

During 2000, when observed mark differences from scripts within an R3 sample tended to be prevalent but small, a possible course of action was to re-mark the work of only those pupils that lay a few marks below each borderline threshold. The underlying logic is that the small marking errors in the remaining scripts would generally not lead to upward level changes.

In agreement with the Lead Chief Marker for KS3 English, it is argued that the decision to re-mark those scripts that fall a few marks above a borderline threshold is as valid as the decision to re-mark those scripts that fall a few marks below. This would be recommended by the measurement ideal (although the threat to the psychological ideal is acknowledged). It is, therefore, recommended that ‘two-way borderlining’ be undertaken where borderlining is deemed appropriate.

However, it should also be noted that borderlining is not a great way to ensure the measurement ideal as there are likely to be at least a small minority of pupils with larger marking errors – whose levels would change post-review – but whose work would not be considered (Cresswell, 1986). Borderline re-marking cuts down on the amount of re-marking required and is likely to identify many of the pupils for whom a re-mark would lead to a level change; however, whether this is sufficient justification for not extending the re-marking to all pupils within a cohort is not clear.

7.6.4.6 Blind re-marking

Lord Mackay of Clashfern, in his evaluation report to Edexcel, made the following comment.

"... I think there must be a risk that, even subconsciously, an experienced Examiner may be influenced by the marks he or she sees have already been assigned. In court proceedings great care is taken to ensure that a judge is not informed, before he or she takes a decision in a case which is contested, whether
or not there has been an offer made by either side in the case. ... Although I can see
the force of the belief that experienced Examiners are robust enough not to be
influenced by the earlier marking, I think it might be fairer if these marks were
not available to the re-marker and the re-marker had to treat the script as
completely unmarked and without any knowledge of what the earlier marks
assigned on the script were, or the reason for them. On the other hand it is
vitally important in my opinion that the re-marker was involved in the earlier
marking of the paper involved and is familiar with the standard used in that
marking although obviously it is necessary that the marker did not mark that
particular script.” (Mackay, 1999, pp.24-5)

While this is an important technical point, it is clear that blind re-marking raises serious
practical problems. Indeed, it could also be argued, in the case of a review, that the task is
not actually to re-mark, per se, but to judge whether the credit originally awarded was
acceptable. As noted in the AQA-N instructions to markers of English:

“in particular, if your assessment of a question differs from that of the original
marker you should consider very carefully whether this represents an error in
original marking or an acceptable difference in marking.”

Admittedly, it is hard to know exactly what this guidance should mean in practice.
Indeed, this is an issue that should be specifically addressed as it should have clear
implications for practice.

Bearing the foregoing concerns in mind, though, blind re-marking is not explicitly
recommended.

7.6.4.7 Miscellaneous procedural issues

Finally, a few miscellaneous procedural issues that arose during 2000 will be addressed.
The first concerns an issue related to borderline checking.

During the initial marking phase, markers revisit scripts that fall just below each of the
specified borderline marks to reassure themselves that their initial marks were
appropriate. This raises the following question: if a teacher finds a clerical error that does
not lead to a level change, but that would take a pupil into the borderline zone, should the
pupil’s work be borderline? Effectively, this would mean re-marking the script.
Although it might seem somewhat excessive, it would seem hard to justify not allowing
this.
Secondly, it was suggested that there might be a case for mark adjustment for R3 reviews. That is, if a sample re-mark for a school revealed that a marker had been consistently harsh or consistently lenient, a decision could be taken to add or subtract a set number of marks from each pupil (thereby avoiding a full re-marking of scripts). This happens for GCSE/A/AS and is generally assumed to lead to a more accurate overall distribution of marks for the schools concerned. However, it should be noted that this does not necessarily imply that each individual pupil will receive more accurate marks. Indeed, recent Australian research has cast doubt on simple mark adjustment with the finding that levels of lenience/harshness often differed, for individual markers, between their first and last marking day (Congdon and McQueen, 2000). Moreover, as national curriculum test scripts are returned to schools, any obvious injustices for individual pupils would be very apparent and this would threaten the psychological and the political ideals.

A possible response is that threats to the psychological ideal could be avoided by only adjusting marks upwards. While this might be true, it would not necessarily support the measurement ideal, nor would it support the political ideal. Indeed, it is easy to imagine a school not complaining to its EMA about individual pupils now being awarded too high a level, but still expressing its concerns in a despairing letter to the TES. In light of the threat to the psychological and political ideals, as well as to the measurement ideal, it is not recommended that mark adjustment be considered for R3 reviews.

The validity of this recommendation might be judged in relation to GCSE/A/AS mark adjustment in future years. As scripts are increasingly returned to schools for scrutiny, the kind of discrepancies discussed above will become increasingly apparent. If, in fact, schools are quite happy to accept the individual injustices that are inevitable with mark adjustment then the recommendation not to adjust for key stage 2 and 3 might be considered less forceful.

7.6.5 Procedures for reporting outcomes of reviews to schools and the NDCA

Procedures for reporting outcomes of review to schools and to the NDCA appeared to function effectively during 2000. However, concern arose as to whether the outcome reports that accompanied rejected requests were received as well as was intended.

7.6.5.1 Reporting outcomes to schools

After the conclusion of a review, a letter is sent to each school noting whether their requested level changes were granted. When level changes are not granted, the EMA also provides a brief explanation of why not in a review outcome report. This is done in
service of the psychological ideal and the educative ideal – to help schools and pupils come to terms with their failed review request(s) by explaining the accepted interpretation(s) of the mark scheme.

Within each EMA, markers, administrative staff and management all devoted a considerable amount of time to processing review outcome reports during 2000. Moreover, there was a strong feeling from the Study 5 participants that the amount of time and effort expended was very worthwhile, as the following comments illustrate:

“I put great value, and I know my chiefs put great value, in their report; because it is a personal request from a school, it means a great deal to the person being reviewed and we treat it as such and so ... particularly if we do not uphold it, we take great pains to explain to them” LCM/M

“Yes, totally justifiable. I think there are several benefits to the review process and one of them is educating – and I don’t mean this in a condescending fashion – but educating people in schools about science, about the tests, about mark schemes, about rigorous marking, and so on.” LCM/S

Unfortunately, there was strong evidence from Study 4 that neither the educative ideal nor the psychological ideal was being achieved. Respondents generally did not agree that they understood the reasons why their review(s) had been rejected, nor did they generally agree that the explanations given addressed the concerns that they had raised. Worse still, not even one-third of respondents, in any sample, indicated that they agreed with the reasons given and at least two-fifths of respondents across samples indicated explicitly that they did not agree. Generally speaking, the review outcome reports appeared not to be educating schools. Indeed, in light of the extent of disagreement with justifications given in the reports, it seems likely that they may have succeeded in further alienating many teachers.

It may well be that trying to educate teachers at this stage of the process is not feasible. Having invested considerable time and effort in the attempt to rectify perceived errors, their psychological defences may be too primed to receive contrary information in an objective manner.

Even if the educative goal was feasible at this stage, though, it is still questionable whether the preparation of individual review outcome reports is the best approach to take. That is, only review-requesting schools are given the benefit of this education and, even then, each school is only informed about issues specifically relating to its own case. If nothing else, it does not seem to be a cost-effective approach.
It is recommended that alternative approaches to educating schools be investigated. For example, the publication of a full report on the review process that noted the frequency of different types of requests and that explicitly discussed the commonly occurring reasons for inappropriate requests (although, this data would not be readily accessible if schools were no longer required to specify which questions they believed to be incorrectly marked). The document could also include a full discussion of questions that most frequently resulted in marking errors.

It is also recommended that a reporting procedure more akin to GCSE/A/AS practice be considered. These reports would not go into detail as to why a review request was rejected. There might, perhaps, be space for one (or two) of a range of general explanations to be ticked.

Then, if schools were insistent that they wanted an explanation for why their request was rejected, they could request a review outcome report as a separate post-review service (for which there would be an appropriate charge). Alternatively, this educative ideal could be achieved more personally through site visits by a local Team Leader (or less personally, through contact by fax, etc.).

Finally, it is recommended that the possibility of reporting review results to named representatives during the summer holiday period be explored.

7.6.5.2 Reporting outcomes to the NDCA

Procedures for reporting review results to the NDCA appeared to function effectively during 2000. Steps had been taken to improve the checks and balances within the system and, particularly, to ensure that data transfers were well monitored. Although no specific concerns were raised about the security of the electronic data transfer, it was felt that the system could be made more robust through, for example, transfer by ISDN link. It is recommended that this option be pursued.

7.6.6 Procedures for dealing with complaints against, or concerns with, the processing of review requests

During 2000, there were no formal procedures for lodging non-review-related complaints against initial marking or for raising concerns with the processing of review requests. Instead, an informal procedure applied through which schools that wanted to complain against the service provided by their EMA were advised to contact the QCA directly. The QCA would only take action on allegations concerning the implementation of procedures (and not on the quality of re-marking).
7.6.6.1 A general complaints procedure

It was suggested on a number of occasions that there ought to be a mechanism through which schools could raise concerns with the quality of initial marking. It was clear that a number of schools felt that they had many scripts that had not been well marked but that would not necessarily have led to review requests. It was felt that there should be a mechanism by which concerns such as these could feed back into the system in order that poor marking and poor markers be identified, at least for developmental purposes. It is recommended that this possibility be explored.

7.6.6.2 A mechanism for appeal

Study 4 asked representatives from schools that had requested R2 or R3 reviews during 2000 whether they believed that there should be a formal appeal process. There was general, although not unanimous, agreement; most notably, over three-quarters of KS3 English R3 respondents agreed. Indeed, two-fifths or more of respondents in each of the three English samples indicated that they would have considered lodging an appeal during 2000, if there had been an opportunity to do so.

Study 5 asked professionals involved with the system for their views. Again, there was a mixed response. While some saw it as a basic human right and a mechanism for ensuring public confidence, it was also suggested that it might transmit inappropriate messages, i.e., that it might encourage schools to complain unduly.

A mechanism for appeal would be supported by each of the principles identified earlier (the measurement, political, psychological and, to a lesser extent, the educative and the accessibility ideal). As such – and in view of the high stakes that are increasingly being attached to test results – it is to be recommended. However, the risk that the system might be overburdened by inappropriate appeals should also be borne in mind. As such, it would be appropriate to attach a deposit/fee to this service (refundable if the appeal was accepted). It would also be important to explain very precisely what would constitute an acceptable reason for lodging an appeal and what criteria would be used for judging appeal cases.

This, in turn, means that QCA would have to consider the scope of the appeal process. In the GCSE/A/AS case, the appeal process extends to concerns over the manner in which papers were set, scripts were marked, grades were awarded and enquiries were conducted. If the remit was this wide then it would have to be directly administered by the QCA.

It is recommended that a formal mechanism for appeal be developed. It might well be appropriate to model this on the three-stage approach operated by the GCSE/A/AS
awarding bodies (EMA first, then QCA). If the first stages of appeal were to be conducted within EMAs, then it is recommended that the appeal process be restricted solely to issues of marking (and not test development, etc.).

If a formal mechanism for appeal is adopted, then it will mean that the process of awarding levels to pupils will continue considerably beyond September of each year. The QCA and DfEE would have to consider how to deal with the issue of finalising and reporting school performance profiles.

7.6.7 Procedures for auditing, monitoring, or evaluating the effectiveness of the review process and for determining ‘customer satisfaction’

The External Marking Agencies were generally less than satisfied with this aspect of the review process and it was clear that there was scope for improvement. In one sense the review system was monitored very closely; this occurred through a Management Report reporting process, through which the EMAs kept the QCA informed about numbers of review requests and the proportions processed by set dates. In other senses, though, there was little in the way of formal monitoring. As discussed earlier, the monitoring of clerical checking and re-marking was not carried out either formally or consistently.

There was also scope for improving general evaluation and reporting practices. For example, with the preparation of reports on the operation of the review system, which included relevant statistics, descriptions of common marker mistakes and descriptions of common inappropriate review requests.

EMA managers felt that more informal discussion and collaboration between EMAs would be valuable and senior markers stressed that they had an important part to play in such developmental processes (often having felt left out during 2000).

While it was felt that there was scope for improving public relations with schools and for closer involvement with them, it was also noted that such exercises were not costed within the external marking contract.

It is recommended that the QCA give further consideration to how procedures for the annual monitoring and evaluation of the review system can be improved. Particularly at this stage of the annual cycle, it is easy to overlook such issues. Yet if public confidence in the review system is to be preserved, it is important that they are not ignored.
7.7 Between the present and the future

The conclusion of the evaluation project is that the structures and procedures of the review system during 2000 were generally effective and fair and operated with sufficient consistency between the three English EMAs. Yet there is considerable scope for improving the technical functioning of the system as well as for improving the definition and documentation of its key principles and practices. Moreover, there is work to be done to improve the accessibility of the system and to ensure that it is perceived as credible.

Part of the remit of the project was to explore differences in principles and practices, not simply between the EMAs, subjects and key stages, but also between the national curriculum perspective and the GCSE/A/AS perspective. While numerous differences between these two perspectives were observed – and while the review system has much to learn from the enquiry-upon-results system – it is not necessarily the case that differences between the two perspectives are indefensible. An evaluation of defensibility ultimately defers to the purposes that an assessment is intended to support. The national curriculum test system and the GCSE/A/AS exam system are sufficiently distinct in purpose to support different assessment models.

On the other hand, there is a clear need to reduce levels of inconsistency within the national curriculum review system. If certain procedural differences between subjects or key stages are deemed to be necessary, they should be given explicit justification. Differences between EMAs, however, should be avoided.

There was not a substantial amount of evidence to indicate that the numbers of review requests are increasing over time. However, it will be important to continue monitoring the prevalence of R2 reviews for English, as increasing trends were most apparent here. Moreover, it is important to realise that the stakes of national curriculum testing continue to increase for schools and this might be expected to impact upon the extent to which schools are prepared to tolerate marking error in the future. In particular, it will be important to consider the potential impact of performance tables at key stage 3.

The preceding discussion and recommendations were intended to provide the QCA with the resources to reflect at length upon the quality of the review system. Indeed, the variety of suggestions offered could support a very different model, perhaps one more similar to the enquiry-upon-results system. The basic procedures underlying such a model are outlined below for illustrative purposes.
7.7.1 An alternative review model

A future system might, like the present system, offer three review services:

R1  a full clerical re-check for individual pupils;
R2  a full re-mark for individual pupils;
R3  the re-marking of an entire cohort.

For R2 reviews, teachers would not be required to specify where they believed marking error to have occurred. Instead, scripts would be re-marked in full.

Neither would teachers be required to specify the nature of perceived marking errors, nor an appropriate sample, for R3 reviews. Instead, they would be required to submit their entire cohort and the EMA would select the sample (although the school would also indicate any pupils that it definitely wished to be re-marked if the R3 was rejected). R3 reviews would be possible for any subject and for either key stage, although teachers should be advised to submit individual R2 review requests unless they believed that more than a specified percentage of their cohort had been awarded incorrect levels.

There would be a fee for each re-marking service that performed a mild deterrent function but that was not prohibitive. Schools would only be invoiced if their review(s) had resulted in rejection. Fees would be returned on a pro-rata basis relating to the success of the outcome.

Schools would submit reviews to the Lead Agency responsible for the subject, rather than to the EMA responsible for marking their pupils’ scripts. The Lead Chief Marker for the subject would be responsible for nationally co-ordinating re-standardisation, re-marking and monitoring. The primary mechanism for the re-standardisation, re-marking and monitoring would be the team marking meeting (which could be conducted either nationally or regionally).

Review re-markers would record only minimal comments for each pupil on an appropriate form. These comments would not be communicated to schools, even in the case of a rejected review. If a school wanted to know exactly why its review was rejected then it would have to complete a form explaining why it believed the script deserved more credit and request a full response from the Lead EMA as an additional service (for which an additional fee would be required).
If a school wished to question the result of a re-mark it would be at liberty to do so but would be required to follow a formal procedure and would have to pay an additional (potentially refundable) fee.

Each year, after the review cycle had been completed, each Lead Agency would be responsible for producing a full report on the general operation of the system. This would include a range of quantitative and qualitative evidence and would be intended to perform an educative function. This would be published and despatched to schools with the return of test scripts the following year, accompanied by a booklet that explained post-testing and post-review services in full.
Section 8 References


