Review of Quality of Marking in Exams in A levels, GCSEs and Other Academic Qualifications

Interim Report

June 2013

Ofqual/13/5287

Contents

1. Introduction

Good marking is a cornerstone of a good exam system. When stakes are high for candidates, and for schools and colleges, everyone needs to have confidence in the grades awarded to candidates and the marking that leads to those grades.

We find that most of the education community, students and the public do have confidence in the quality of marking of exams in A levels and GCSEs (Ipsos MORI, 2013), and yet preliminary grades and marks are increasingly contested. What is more, a significant (and growing) minority of teachers and head teachers tell us that they do not believe that marking has been good enough in recent years, especially in GCSEs. Confidence in marking fell noticeably last year because of concerns about GCSE English, but, even recognising the unusual circumstances, the recent trends are troubling.

We are, therefore, reviewing the quality of marking of A levels, GCSEs and other academic qualifications (referred to collectively as general qualifications). Our review is limited to the marking of external exams, and does not cover controlled assessment or any other internal assessment. We will report outcomes of the review in three stages.

There have been significant developments in marking in the last decade, with an accompanying body of research on those developments. In this, our first report on marking, we set out how marking works today and comment on the most significant developments. Alongside this report, we are also publishing a literature review exploring some of the relevant research. This *Review of Literature on Marking Reliability Research* is available on our website.[1]

In this report, we go on to explore some common criticisms of the marking system, and identify those aspects of marking where we are doing further work. We also summarise the preliminary results from a large-scale survey of examiners, which we have undertaken as part of our review. Contrary to some beliefs, we find that examiners are knowledgeable, nearly always holding a degree in their subject; that nearly all have many years of experience teaching their subject, often as head of department; and that most examiners have been marking for many years. If this were more widely known, it should promote greater public confidence.

Our second report will be published in early August and will focus on the arrangements for challenging grades and marks, including the enquiries about results (EAR) and appeals processes. Head teachers and teachers tell us they are

[1] www.ofqual.gov.uk/files/2013-06-07-nfer-a-review-of-literature-on-marking-reliability-research.pdf

particularly concerned about how these processes work, and are not always confident in the outcomes. In our final report, which we will publish in the autumn, we will detail the results of our further work on the quality of marking of exams in England, and make final recommendations for the marking system.

When we refer to “quality of marking” we mean both the accuracy and reliability of marking. This is to say that candidates should receive marks as close to their correct, “true” scores as is possible, and that this should be the case no matter who marks their work. Evaluating quality of marking is not straightforward: there is no single accepted way of measuring marking quality and few common metrics are available. Nonetheless, there are characteristics that we expect to see in a healthy marking system. For instance, we expect exam boards to have robust systems and controls to promote good marking, to prevent poor marking, and to identify and remedy poor marking when it happens. We expect exams to be marked by examiners with the right skills and experience. And we expect any review of a mark through the EAR or appeals process to be dealt with consistently, fairly, transparently and promptly.

There are limitations to any exam system. A mark is a human judgement of a candidate’s work and is only ever an approximation of the candidate’s “true” score. If we are to have valid assessments that measure the right skills, knowledge and abilities in the right way, marks can never be totally reliable. Multiple-choice responses can be marked with precision, but long-answer questions will always leave scope for differences of opinion between equally qualified and skilful examiners, as there is no right answer. As key qualifications are reformed, we anticipate more assessment each summer, more assessment by exam and more complex long-answer questions in some subjects. Anticipating these changes, the quality of marking needs to be as good as it can be.

As we set out in this report, the biggest factor influencing reliability of marking is the design of the assessment - the style and quality of the questions and the quality of the accompanying mark schemes. These are matters we intend to improve, as qualifications are reformed. However, marking is not just a matter for the regulator, or exam boards. Some 51,000 individuals (known as examiners) mark exam scripts each year. They each play their part in the wider public institution of awarding key qualifications, and we recognise and value the important contribution they make.

We hope that this report will enable you to understand how marking works today. We will soon report on EARs, where there are specific concerns, and we look forward to reporting finally on marking in the autumn, once we have completed the further work we are doing to see to what extent marking can be improved.

2. How marking works

GCSEs, AS and A levels are the main academic qualifications taken by candidates in England. In summer 2012, 1.27 million candidates took GCSEs in 48 subject areas[2] whilst over half a million candidates took A levels (or AS) in 36 subject areas[3]. A much smaller number of candidates took level 1 and 2 certificates (known as IGCSEs), the Pre-U Diploma (or Pre-U Certificates) and the International Baccalaureate (IB) Diploma.

Almost all general qualifications include externally set and marked exams. In modular qualifications (which are split into different units), candidates are examined at the end of each unit. In linear qualifications, they are examined at the end of a course. After candidates sit an external exam, the completed answer booklets (candidate scripts) are sent to exam boards for marking. Scripts are marked by external examiners who score each question using a set mark scheme.

Many qualifications also include an internally assessed element of coursework or controlled assessment, which is either set by teachers within a school or college (within parameters defined by exam boards) or set by exam boards. This work is marked by teachers using marking criteria provided by exam boards. Exam boards moderate samples of candidates' work to check that marking has been carried out correctly by teachers.

The ratio of external to internal assessment varies. In GCSEs, internal assessment can be 25 per cent or 60 per cent of the total, or the qualification can be entirely externally assessed. We set this ratio of internal to external assessment through subject criteria. At A level, the proportion of internal assessment is more variable. Some subjects, such as art and design, are entirely internally assessed. In contrast, no more than 20 per cent of A level in Maths can be internally assessed. Example assessment profiles are shown in appendix A. The marking of internal and external assessments follow two quite distinct processes. In this review of quality of marking, we consider only external exams marked by exam boards.

2.1 Marking of external exams

In summer 2012, over 15 million GCSE and A level (including AS) external exams were taken by candidates in England, Wales and Northern Ireland, resulting in the issue of around 7.5 million GCSE and A level results. The breakdown of exams taken from each exam board is provided in the table below.

[2] 1,270,118 candidates were entered for GCSE qualifications in England, Wales and Northern Ireland (Joint Council for Qualifications, 2012).

[3] 334,210 candidates were entered for A level qualifications and 507,388 for AS qualifications in England, Wales and Northern Ireland. Please note that these figures cannot be combined to calculate a total A level figure (Joint Council for Qualifications, 2012).

Exam board	GCSE external exam scripts marked	A level external exam scripts marked
AQA	4,259,000	1,683,000
CCEA	264,000	115, 000[4]
OCR	2,871,000	1,026,000
Pearson Edexcel	2,638,000	1,143,000
WJEC	1,170,000	284,000
Total	11,202,000	4,251,000

In the same exam series, 531,000 scripts were marked for other general qualifications, including level 1 and 2 certificates (known as IGCSEs), IB Diplomas and Pre-U Certificates.

Other general qualification	External exam scripts marked
Level 1 and 2 certificates (known as IGCSEs)	443,000
IB Diploma	78,000[5]
Pre-U	10,000

The time pressures on exam boards to process this volume of scripts is great. In 2013, GCSE and A level exams began on 13th May and run until the last week in June[6]. Results are issued to candidates on 15th and 22nd August respectively (Joint Council for Qualifications, 2012). Certain A level results, therefore, need to be processed within seven weeks of candidates taking an exam. The IB Diploma timescales are tighter still, with all scripts processed within six to nine weeks of the exam (International Baccalaureate Organization (IBO), 2012).

The management of this high-volume process is complex. Examiners are a highly geographically distributed workforce who mark scripts from home. For GCSEs and A levels (including AS), examiners are generally based in the UK, although Pearson Edexcel does also have one general marking facility in Melbourne, Australia. For international qualifications such as the IB Diploma, examiners are spread across the globe. In both instances, monitoring of marking must, therefore, take place remotely.

[4] Includes Applied GCE

[5] The International Baccalaureate Organization externally assessed 1,437,853 scripts in May 2012, of which approximately 78,000 were from UK schools.

[6] The final A level exam will be held on 24th June and the final GCSE on 26th June.

Examiners do not work for exam boards on a full-time basis. They are contracted to work on a single exam series and usually organise their marking work flexibly around other employment. This can present its own difficulties, with 43 per cent of examiners telling us that fitting marking in around their main job can be “very” or “somewhat” challenging[7].

Whilst the introduction of new technologies has helped to lessen some of these challenges, the logistics of this process still have to be carefully managed. As such, exam boards rely on tightly controlled marking processes and quality controls. We are studying all aspects of these arrangements to assess where there might be room for improvement.

2.2 Who marks exam scripts?

The literature review shows that experienced examiners are crucial for good marking of exams requiring complex, extended answers (Tisi et al., 2013a). Few studies have tried to pinpoint exactly which aspects of experience are most important – whether it is examiners’ subject knowledge, teaching experience or marking experience. However, some studies show that subject knowledge and teaching experience are the most critical to quality of marking, more so than previous examining experience (Meadows and Billington, 2007).

There is also evidence to show that as questions become less complex, examiner experience is less important. For simple, highly constrained questions, general markers with no subject knowledge or examining experience can mark just as reliably as very experienced examiners (Tisi et al., 2013b; Meadows and Billington, 2005a).

In summer 2012, over 51,000[8] examiners marked GCSEs, A levels (including AS) and other academic qualifications[9]. Almost all of these were examiners with considerable teaching experience and subject knowledge. A minority were general markers who mark simple, highly constrained questions with clearly defined answers. General markers are generally used sparingly by exam boards. For example, in 2012, AQA used over 17,000 markers, of whom around 100 were general markers, to mark GCSEs and A levels.

Until recently there was no data on the profile of examiners across the system. To address this, we surveyed examiners during April and May 2013, attracting over

[7] Figures represent initial results taken from our *Survey of Examiners 2013*.

[8] This does not include data from the Council for the Curriculum, Examinations and Assessment.

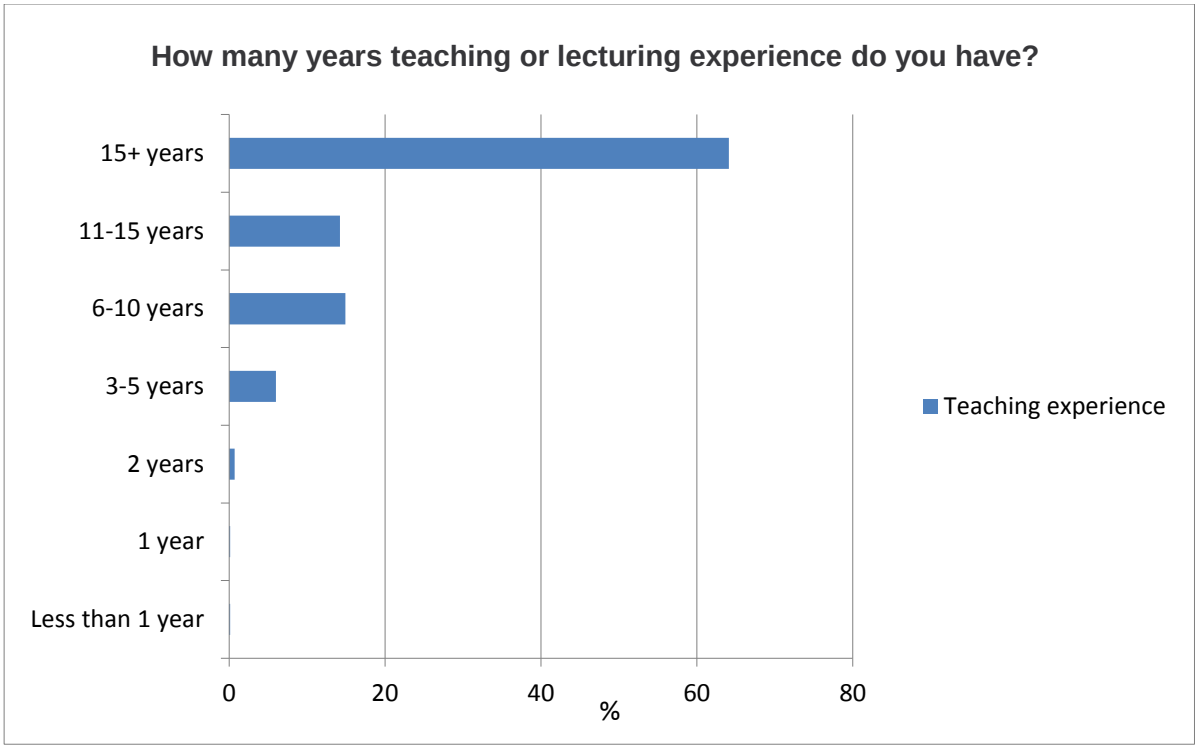
[9] This figure will include some double counting, as some examiners mark for more than one exam board.

10,000 responses[10] – at least one in five of the workforce. Some provisional findings from our survey are discussed below. We cannot compare these statistics to any existing data on examiners, so it is difficult to say whether these figures are completely representative of the population. However, there is no known reason why these figures would not be representative of the workforce, particularly given the size of the response. Our initial findings show that examiners are knowledgeable, nearly always holding a degree in their subject; that nearly all have many years of experience teaching their subject, often as head of department; and that most examiners have been marking for many years.

We know that nearly all examiners are, or have been, teachers. Typically, exam boards only recruit examiners if they have some degree of teaching experience. Our survey of examiners found that over 99 per cent of respondents have teaching experience. Just under two thirds (62 per cent) are currently teaching, with 38 per cent former teachers or lecturers. Most examiners teach the same specifications that they examine. Learning more about these specifications was cited as the single most important motivation for becoming an examiner.

Most examiners are also experienced teachers. Almost two thirds of the examiners surveyed (64 per cent) have over 15 years' teaching experience, with 93 per cent having six or more years' experience. Only 0.2 per cent have less than two years' teaching experience.

[10] Cambridge International Examinations carried out its own survey of examiners. This will be analysed alongside our survey data in our final report.



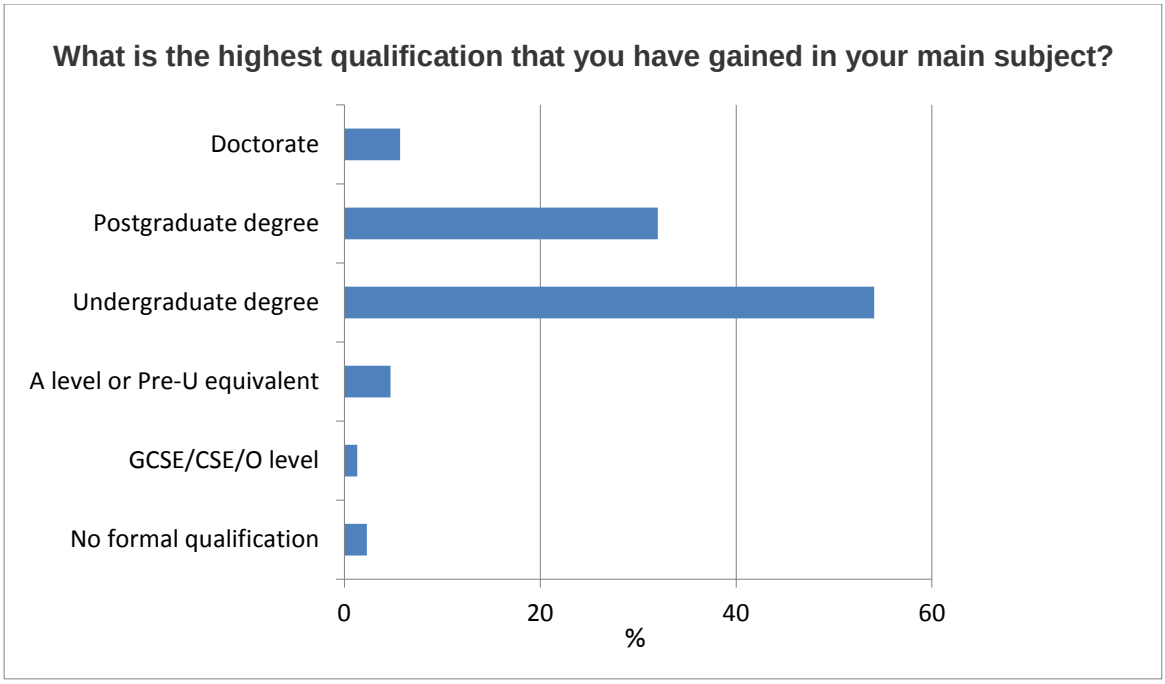
Effective base: 10,151 examiners who were, or had been, teachers or lecturers (April to May 2013).

Examiners are also often quite senior. Over a third (35 per cent) of respondents are, or have been, a head of department, and 4 per cent have been a head of year. Seven per cent are, or have been, a head teacher or a deputy or assistant head teacher.

Most examiners have been involved in marking for some time. Almost half of the respondents (47 per cent) had examined for over ten years, with around seven in ten (69 per cent) examining for more than five years. Thirteen per cent of the examiners surveyed had less than three years of marking experience.

As well as teaching experience, we also know that most examiners have considerable subject expertise. More than nine in ten examiners (92 per cent) have a degree (postgraduate or undergraduate) or a doctorate in their main subject.

Just 2 per cent of examiners have no formal qualification in their main subject. They are often experienced examiners who work for a range of exam boards. Usually, these examiners mark newer subjects such as ICT, citizenship or media studies, or subjects that draw on a range of disciplines, such as general studies. Others mark modern foreign languages and may include those for whom a modern foreign language is a home language.

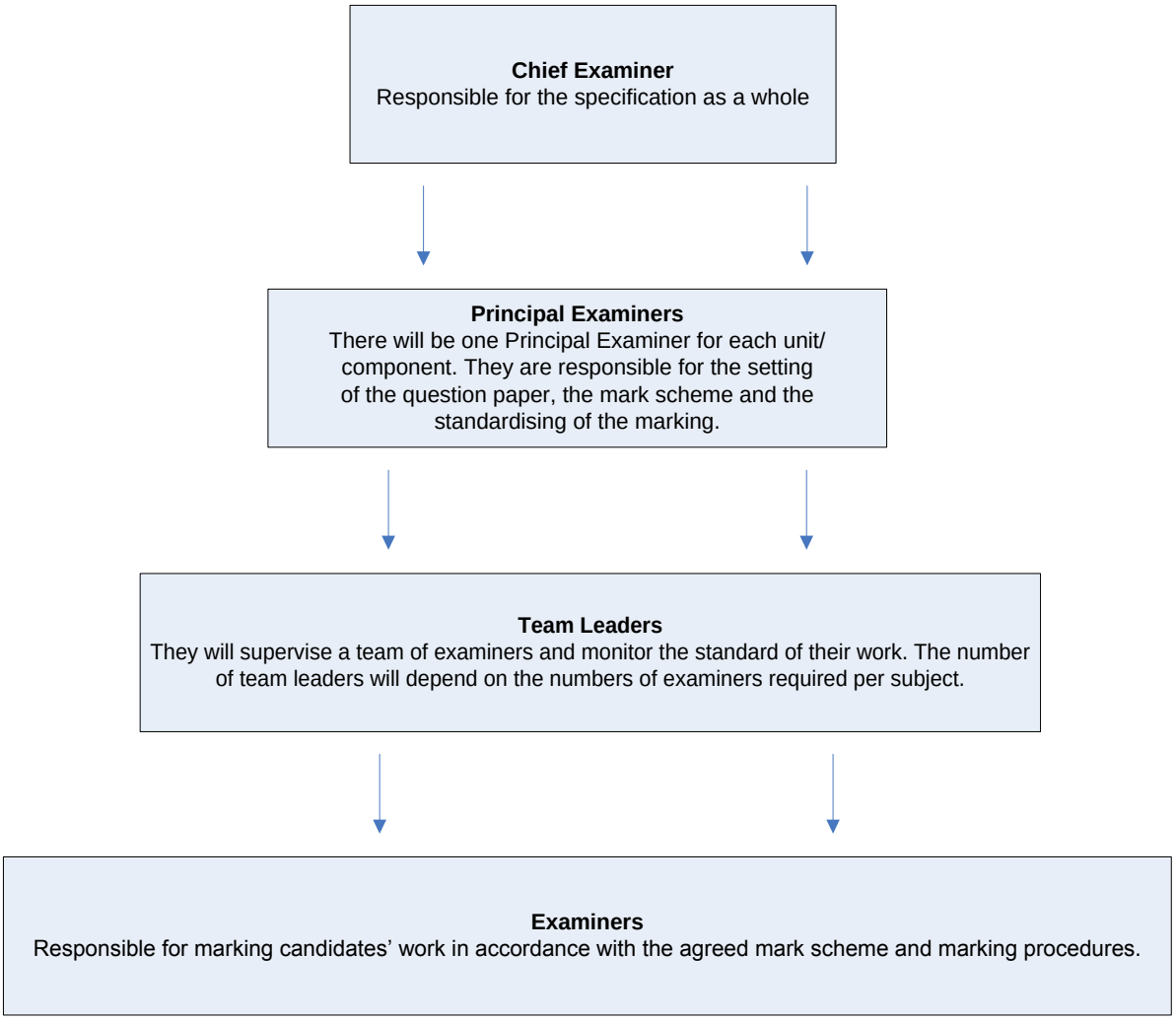


Effective base: 10,205 examiners (April to May 2013).

2.3 How are exams marked?

Each qualification is marked by a team of examiners, led by a chief examiner who is responsible for the qualification. The chief examiner reports to the chair of examiners, who is responsible to the awarding organisation for maintaining standards across different specifications in a subject within a qualification and from year to year. The chief examiner is supported by principal examiners who are each responsible for a particular unit, or module. The chief examiner and principal examiners are responsible for developing the question papers and their mark schemes, supported by a team of revisers and scruntineers (Ofqual, 2011a). The principal examiners train their examining team in how to apply the mark schemes.

Depending on the size of the qualification, examiners are typically organised into teams of anywhere between five and ten members. These examiners are monitored by a team leader who is in turn monitored by the principal examiner.



2.3.1 Mark schemes

Examiners score candidates' work by applying a mark scheme. Mark schemes are a set of criteria used to judge how well a candidate has performed on each task or question. They lay down the marking standard and are designed at the same time as a question paper is developed. The quality of a mark scheme is central to an examiner's ability to mark well. In their review of marking reliability in 2005, Meadows and Billington found that "an unsatisfactory mark scheme can be the principal source of unreliable marking" (Meadows and Billington, 2005b, p. 42).

Mark schemes tend to fall into three broad categories. Objective mark schemes are used for questions where there is an unambiguous correct answer and detail precisely the only acceptable answer. Points-based mark schemes are generally used for structured questions requiring no more than one or two paragraphs in response. Marking usually involves counting up the number of creditworthy points made by candidates in their response.

More generic levels-based mark schemes are used for unstructured questions requiring longer answers. These mark schemes describe a number of levels of response, each of which is associated with a band of one or more marks. Examiners apply a principle of best fit when deciding the mark for a response (Bramley, 2008). Levels-based mark schemes can be holistic, where examiners give an overall judgement of performance, or analytic, separating the different aspects of candidates' performance and providing level descriptors (and mark bands) for each aspect. Levels-based mark schemes are the most subjective, particularly if they are holistic in nature. For more information about mark schemes, see appendix B.

2.3.2 Types of marking

Depending on the qualification or the subject, candidate scripts are generally marked traditionally (using pen and paper) or on-screen (electronically) by examiners.

There is a third distinct marking type, which accounts for 1 per cent of all marking of GCSEs, A levels and other academic exams in England. This is automated marking, used for multiple-choice exam papers, predominantly in science GCSEs. Automated marking does not use human examiners; instead marks are allocated using optical mark recognition software. Given the low levels of automated marking, the discussion below focuses on the two main types of marking: traditional and on-screen.

Until relatively recently, all marking was carried out traditionally using pen and paper. In traditional marking, batches of scripts are physically sent to examiners. Examiners generally mark their batches at home and then return their scripts to the exam board for checking. The task of sending scripts to examiners adds additional time and cost to the marking process.

On-screen marking is a relatively recent development, first introduced for general qualifications by Pearson Edexcel in 2003. In on-screen marking, candidate scripts are scanned into digital format and sent to examiners for marking on a computer screen, via a secure system. Since its development, on-screen marking has grown rapidly and is now used by all exam boards to some degree. On-screen marking can speed up administrative aspects of the marking process and it eliminates the need to send candidate scripts around the country.

On-screen marking is now the main type of marking used in general qualifications. In summer 2012, around two thirds (66 per cent) of all scripts[11] were marked on screen. GCSEs and level 1 and 2 certificates (known as IGCSEs) were most likely to be marked electronically (around 68 per cent and 79 per cent respectively). This was followed by A levels (around 60 per cent) and the IB Diploma (61 per cent). In contrast, Pre-U Certificates are entirely traditionally marked.

[11] For GCSEs, A levels and other academic exams taken in England, Wales and Northern Ireland.

Some exam boards (such as Pearson Edexcel) mark almost all their scripts on screen, whilst others (such as the Council for the Curriculum, Examinations and Assessment (CCEA) and WJEC) mark a minority of scripts in this way. The breakdown for each exam board is shown in the table below.

	Percentage of scripts marked online in summer 2012[12]	**Percentage of scripts marked traditionally in summer 2012**
Pearson Edexcel	88%	12%
OCR	79%	21%
IBO[13]	61%	39%
AQA	60%	40%
CIE	32%	68%
CCEA	15%	85%
WJEC	13%	87%

As well as its logistical benefits, on-screen marking should improve marking reliability by enabling more frequent and flexible monitoring of examiners by exam boards. Senior examiners review their team's marking almost in real time, ensuring that inconsistent or inaccurate marking is detected early. Examiners marking on screen input their marks directly into the system. This also reduces the likelihood of clerical errors associated with the incorrect addition or recording of marks.

We can see the logistical benefits of on-screen marking, as well as the real opportunities that it brings for real-time monitoring of marking. However, we know that online systems can introduce new sources of clerical errors. There have been a small number of occasions where systems have not correctly added the marks entered by examiners. There have also been instances where some pages of an answer booklet have not been scanned or viewed online by examiners. This can be more of an issue for essay questions that use generic answer booklets, as it is not always clear exactly where candidates have written their answers. These clerical errors do not necessarily represent poor-quality marking in the traditional sense, but they are a form of marking error. Furthermore, they can have a significant impact on student, parent and school confidence.

[12] Data excludes any scripts that are marked using automated marking. Automated marking made up 1 per cent of all marking in summer 2012.

[13] Includes externally-assessed coursework components

On-screen marking opens up possibilities for changes to established marking processes, which have the potential to improve marking reliability. One of the most significant of these changes is the move to item-level marking, where a scanned script is split up into individual questions (or groups of related questions), which are marked by different examiners. Examiners are able to mark large batches of a particular item, allowing them to become deeply familiar with the mark scheme for that specific question as well as a full range of candidate answers. This is a departure from the traditional approach of one examiner marking a whole candidate script. Where exam boards use on-screen marking, many also use item-level marking for these scripts, although this is not always the case. AQA, Pearson Edexcel and WJEC all use item-level marking for their on-screen scripts, whereas OCR, IBO, Cambridge International Examinations (CIE) and CCEA use whole-script marking.

We know that there are differing views about item-level marking in the education sector. However, assessment research generally supports item-level marking and suggests that, at least in theory, it has the potential to improve the accuracy of marking in exams by reducing the effects of biases caused by the rest of the question paper (the “halo” effect) and by removing the influence of a single examiner on an exam script (Pinot de Moira, 2011b; Spear, 1996a). For a more detailed discussion of item-level marking, see section 4.

In summer 2012, just over half of the exam scripts were marked as a whole (54 per cent), rather than split into items (45 per cent). As we might expect, item-level marking was most prevalent amongst qualifications where on-screen marking was highest. In GCSEs and level 1 and 2 certificates (known as IGCSEs), item-level marking made up around 46 per cent and 71 per cent of marking respectively. At the other end of the spectrum were the IB Diploma and Pre-U Certificate, where candidate papers were only ever marked as a whole script. Levels of item-level marking also vary significantly by exam board. Currently, IBO, OCR, CIE and CCEA do not use item-level marking, whilst for Pearson Edexcel, 88 per cent of scripts are marked at item level.

Future developments in marking

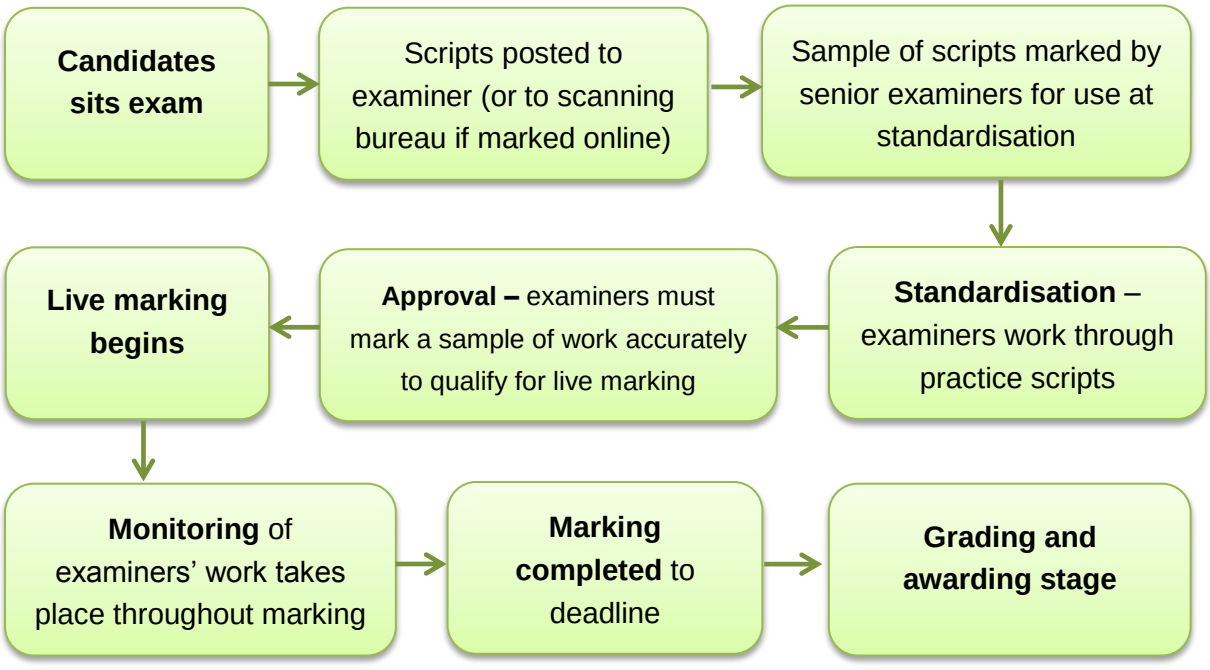
The data above presents a snapshot of marking as it was during summer 2012. We know that all exam boards intend to extend on-screen marking. As part of this shift, some (but not all) exam boards intend to increase their use of item-level marking. We can, therefore, expect to see a further decrease in traditional marking over the coming years. As we discuss in section 4, assessment research shows that both of these developments have good potential to improve marking reliability further, however, we believe that there is a need for further empirical evidence to show how they are working in practice for key qualifications.

GCSE reform will also influence the marking approaches used by exam boards. GCSE reform is likely to increase the number of more stretching, extended tasks and reduce the number of highly structured questions. This places a greater emphasis on the role of the experienced examiner, and so the use of general markers is likely to remain low.

2.3.3 Marking in practice

The details of marking arrangements vary by exam board, but the general principles remain the same. This means that the journey of a script is the same across all general qualifications. Where there are significant differences in the marking process, this is driven by the type of marking used – in other words whether an exam is marked on-screen or traditionally.

The general marking process used by all exam boards is shown below. This shows the journey of a script from the moment the candidate completes the paper up to the point when the script has been fully marked and a grade is ready to be awarded. This diagram is a generalisation of a complex process and may not apply to all scripts. For a more detailed description of the marking process (including how it varies between on-screen and traditional marking), see appendix C.



After awarding, further quality checks may take place for a range of reasons. For example, additional checks may be performed on any scripts where there are doubts about the performance of an examiner, or where the performance of a candidate or school is significantly different from expectations.

Three of the stages in the diagram above are particularly critical for ensuring marking quality. They are standardisation, approval and monitoring.

Standardisation

Good marking depends upon the examiners' shared understanding of the mark scheme, and the consistency of its application (Chamberlain and Taylor, 2010). Before examiners start marking, all go through a standardisation process. This aims to ensure that examiners are fully competent in applying the mark scheme consistently before they begin marking. As this process is specific to each exam paper, standardisation takes place for each examiner in every exam series. Standardisation takes place after an exam has been taken, typically within a week of the exam date.

During standardisation, examiners practise marking completed exam scripts using a mark scheme to build up an understanding of the marking standard that they must apply. Standardisation is either carried out in face-to-face meetings (generally used for traditional marking) or online (generally used for on-screen marking). There is little evidence to show the effectiveness of different methods of standardisation, and the research that has been carried out gives a mixed picture.

Until recently, standardisation was always carried out face to face. These face-to-face meetings are chaired by principal examiners, who may be supported by team leaders. More and more often, however, standardisation is being delivered online. In online standardisation, examiners work through a series of candidate scripts on-screen. These scripts are clearly annotated to show exactly how the principal examiner has applied the mark scheme. Some exam boards, such as Pearson Edexcel, also use interactive approaches, using web platforms to host online standardisation meetings. Here participants can interact in remote discussions, simulating a face-to-face meeting.

This move from face-to-face to online standardisation has not always been well-received by examiners and has led to a turnover in some examining teams. Preliminary results from our survey of examiners indicate that some examiners feel very strongly about the loss of face-to-face standardisation. Critics believe that online standardisation does not facilitate the same depth of discussion and interrogation of mark schemes as face-to-face meetings do. Removing face-to-face meetings means that there is no community of practice to develop a comprehensive shared understanding of how a mark scheme works. This community of practice may be important in more subjective disciplines, such as English and history, where the consistent application of a mark scheme is more difficult.

In their 2005 literature review on marking reliability, Meadows and Billington cited research that highlights the importance of examiner meetings as a means of

internalising a mark scheme and they suggested that these meetings could not be replaced by any amount of detailed, written instructions (Meadows and Billington, 2005c). Aside from any possible impact on marking quality, the loss of face-to-face contact appears to have an impact on examiners' feelings of engagement with the marking system, leading to a certain sense of disconnect. Web conferencing facilities attempt to simulate this community of practice remotely. However, there is little evidence as to the effectiveness of this. More generally, there is no evidence that suggests whether one type of online standardisation is more successful than the other.

There is no doubt that online standardisation has logistical advantages for exam boards. It is quicker and cheaper and it caters well for a geographically distributed workforce. In some international qualifications with a global examining team, such as the IB Diploma, online standardisation is vital in ensuring consistency of examiners where a face-to-face meeting involving all examiners is not always practical. For all exam boards, however, it enables the retention of examiners who may not have been available to attend a face-to-face meeting on a certain day. It also has other benefits. Whilst traditional standardisation may be delivered by a number of team leaders in large examining teams, online standardisation provides one consistent message directly from the principal examiner. There is, therefore, no risk of a dilution in the marking standard through different team leaders. In online standardisation, training materials are always available on the online environment, and so examiners can return to them as often as they like to check their understanding. A study published by AQA showed that even in a subject such as history, online standardisation can be as robust as face-to-face approaches (Chamberlain, 2010a).

As on-screen marking increases, the use of online standardisation looks to increase further. Given the debate and the conflicting research findings, more research is needed to evidence the benefits and drawbacks of this in practice.

Approval

After the standardisation phase comes an approval phase. This is the point at which a judgement is made as to whether examiners are ready to graduate to live marking. In order to qualify, examiners must independently mark a sample of scripts or questions either on-screen or on paper. Their marking is reviewed by a senior examiner[14] who makes sure that the work is up to the required standard. This is usually measured through the application of a marking tolerance, which compares

[14] A senior examiner may be a team leader or principal examiner.

the mark given by an examiner to the mark that a senior examiner would give to the same work. The exact marking tolerances used vary by exam board[15].

The use of tolerances recognises that there can be legitimate differences in professional judgement between examiners, particularly in certain subjects or question types. Whereas in history or philosophy an examiner might be expected to mark within 5 per cent of the senior examiner's mark[16], in maths there might be no tolerance at all. In a qualifications market we can accept variation and innovation in practice. However, there are instances in which consistency across exam boards is desirable. We will consider whether consistency across exam boards is desirable for the setting of approval tolerances.

When an examiner has demonstrated that they can apply the mark scheme correctly, they are cleared to begin live marking. If they do not succeed, they are given further training and a second chance to qualify. Examiners who do not meet the required standard at this point are prevented from marking specific questions (if they are marking at item-level) or whole scripts (if the marking is at whole-script level). In on-screen marking, this qualification process is extremely quick; examiners can be cleared to mark almost immediately. In traditional marking, this process can take several days.

For most exam boards this approval process happens once at the start of the marking window. However, at AQA, examiners are required to qualify for live marking each time they log into the on-screen marking system.

Monitoring

Throughout the exam marking period, a sample of examiners' work is checked by senior examiners to ensure that they continue to apply the mark scheme accurately and consistently. One of the real benefits of on-screen marking is that it enables continuous, real-time monitoring; examiner marking is always visible to senior examiners via the online system. The most sophisticated systems allow exam boards to monitor when examiners are marking and the speed at which they mark.

In on-screen marking, multiple types of monitoring can be used. Typically, exam boards plant "seed" scripts (or items) in each examiner's batch of marking, usually at a rate of at least 5 per cent (and up to 10 per cent in the case of IBO and CCEA). These "seeds" have already been given a definitive mark by the senior examining

[15] Tolerances will also vary depending on whether marking is carried out by whole script or at item level.

[16] At whole script or item level.

team and examiners must mark within the tolerance of this mark. Seeding is purely used as a tool to check the accuracy of examiners' marking – it is not a form of double marking. The senior examiner's original mark is the mark that candidates' work used as a seed will receive. Again, the use of tolerances varies across exam boards. We will consider the suitability of this in our final report.

In on-screen marking, other monitoring can supplement the use of seed items. In many exam boards, senior examiners also spot check (or "back read") samples of examiners' work. In these instances the senior examiner is able to view the marks and annotations of the original examiner. Their re-mark is, therefore, not independent of the first mark. Research has demonstrated that there is an improved likelihood of detecting unreliable marking if the second marker is unable to view the marks of the original marker (Tisi et al., 2013c).

Finally, AQA and WJEC also use double marking of a sample of questions in more subjective disciplines as a third way of monitoring marking reliability. If the marks given by the two examiners are out of tolerance from each other, a senior marker will decide what mark should be given to the candidate's work, and a penalty mark will be given to one (or both) examiners.

Once an unacceptable level of inaccurate or inconsistent marking has been identified through any of the methods above, examiners are stopped temporarily. They are given additional support until the exam board is satisfied that they can mark in line with the common, standardised approach. If this is not the case, they are not allowed to continue marking scripts (or specific questions, if item-level marking). If necessary, any work that they have completed will be re-marked.

Such close monitoring of examiners is not possible with traditional paper-based marking. In traditional marking, examiners send senior examiners samples of their marking at two or three agreed points during the exam marking period. The exact details of the sampling process vary for the different exam boards, but, in all cases, the sample should cover a good range of candidate performance and answer types. In general, the senior examiner re-marks 10 to 25 of the sample scripts to ensure that they are consistently within tolerance. This sample of re-marked scripts is used to make decisions about whether the examiner's marks need to be adjusted (if adjustments are used by the exam board) or included in a marking review process, or whether the examiner's allocation needs partial or total re-marking. Marking adjustments are not required in on-screen marking where examiners can be stopped and corrected in real time.

2.4 What happens after candidate work is marked?

The marking process produces a raw mark for each candidate script. This is the total of the marks given for each question on the script. With on-screen marking, this mark

is automatically calculated by the online system. In traditional marking, examiners add up the marks for each question or enter them onto an online system. This is then checked by the exam board. There have been some occasions in the past when these manual marking checks have not picked up errors in the addition or transcription of marks (Ofqual, 2012a).

Once a set of raw marks is ready for each component of a qualification, exam boards can set grade boundaries. This process – known as awarding – is separate from marking and is not considered as part of this review of marking.

After grade boundaries are set, candidates' raw marks can be converted into grades. For modular (and some linear) qualifications, raw marks are converted using the uniform marks scale (UMS). Uniform marks from each unit are combined to give an overall qualification grade. For more information about UMS marks, see appendix D.

For more information about awarding and grading processes, see our website.[17]

[17] www.ofqual.gov.uk/help-and-advice/about-marking-and-grading

3. The challenges facing a marking system

In a high-stakes exam system, it is essential that exams are marked as accurately and reliably as possible. However, we must be clear about what a marking system can ever reasonably deliver. As with many measurement tools, any assessment is likely to include some element of unreliability in its results. Whilst it is not possible to remove all unreliability that exists within marking, we can ensure that marking is as good as it can be in the context of our exam system.

3.1 Validity and reliability of assessment

A high-quality exam system must test candidates in a valid and reliable way. That is to say exams must measure what they are intended to measure, and they must do this in a consistent way. Without either one of these features any assessment is flawed.

Validity is a measure of whether exam results are a good measure of what a student has learned. It ensures that we are testing the right knowledge and skills in the most appropriate way. Achieving validity is the single most important aim of an assessment. Validity is also underpinned by reliability. In simple terms, reliability refers to the repeatability of an assessment and the extent to which a mark is free from any random or systematic error (Nunnally and Bernstein, 1994). We provide a straightforward definition of reliability as part of our reliability programme:

> "Reliability" in the technical context means how consistent the results of qualifications and assessments would be if the assessment procedure was replicated – in other words, satisfying questions such as whether a student would have received the same result if he or she happened to take a different version of the exam, took the test on a different day, or if a different examiner had marked the paper[18].

The reliability of an exam reduces when any type of variation (or error) is introduced into the process. Marking is just one source of possible variation. However, the reliability of marking may be influenced by multiple factors. In their 2005 review of marking reliability, Meadows and Billington found that the single most important factor affecting marking reliability is assessment design – the style and quality of individual exam questions and the mark schemes used (Meadows and Billington, 2005d). This has been reinforced in a number of studies since (Tisi et al., 2013d).

Tightly defined questions with unambiguous answers can be marked much more accurately and reliably than extended-answer questions. It is much easier for

[18] http://www2.ofqual.gov.uk/standards/reliability/

examiners to identify the correct answer to a multiple-choice question than it is to judge the quality of an essay response, for example. As questions become less constrained and more complex, it is harder to determine exactly how good a response is. It also becomes a more subjective judgement, and lower levels of marker agreement on essay questions may be a result of legitimate differences in opinion between equally qualified examiners (Tisi et al., 2013e).

Whilst tightly constrained, short-answer questions will result in the higher reliability of an exam, they are not always a valid means of assessing certain knowledge and skills. Our international research shows that well-constructed multiple-choice questions can be extremely effective in assessing certain knowledge and skills. However, in some subjects the use of high-mark questions with complex, extended responses is an important aspect of validity. Here, an education system may accept the lower levels of reliability where we believe the question type to be essential in assessing certain knowledge and skills. However, if levels of reliability become too low, results are not a consistent measure of candidate performance and the assessment becomes meaningless.

In March 2013, the Secretary of State wrote to us setting out the government's policy on reforms to GCSE qualifications. One aspect of these reforms is to increase the demand of GCSEs through a focus on more stretching tasks and fewer bite-sized and overly structured questions (Gove, 2013). This marks a shift in the balance away from reliability and towards validity.

When making judgements about quality of marking of GCSEs, A levels and other academic qualifications we must, therefore, accept that the exam system will never be able to deliver absolute reliability if we are to measure the right skills, knowledge and abilities in the right way. It does, however, need to be reliable enough to ensure that exam results can be used for their various high-stakes purposes, including accountability. We will concentrate on identifying those improvements that can be made to optimise reliability whilst protecting assessment validity.

3.2 Public confidence in marking

Another challenge for a marking system is maintaining a level of public confidence that exams are marked accurately. As part of this review, we are gathering stakeholder perceptions of the marking process. These perceptions can help us to understand what drives public confidence, and they may identify specific strengths or failings of the system. It is important to note, however, that these perceptions will be based on specific individual experiences and they may not always reflect the wider reality of the marking system in England.

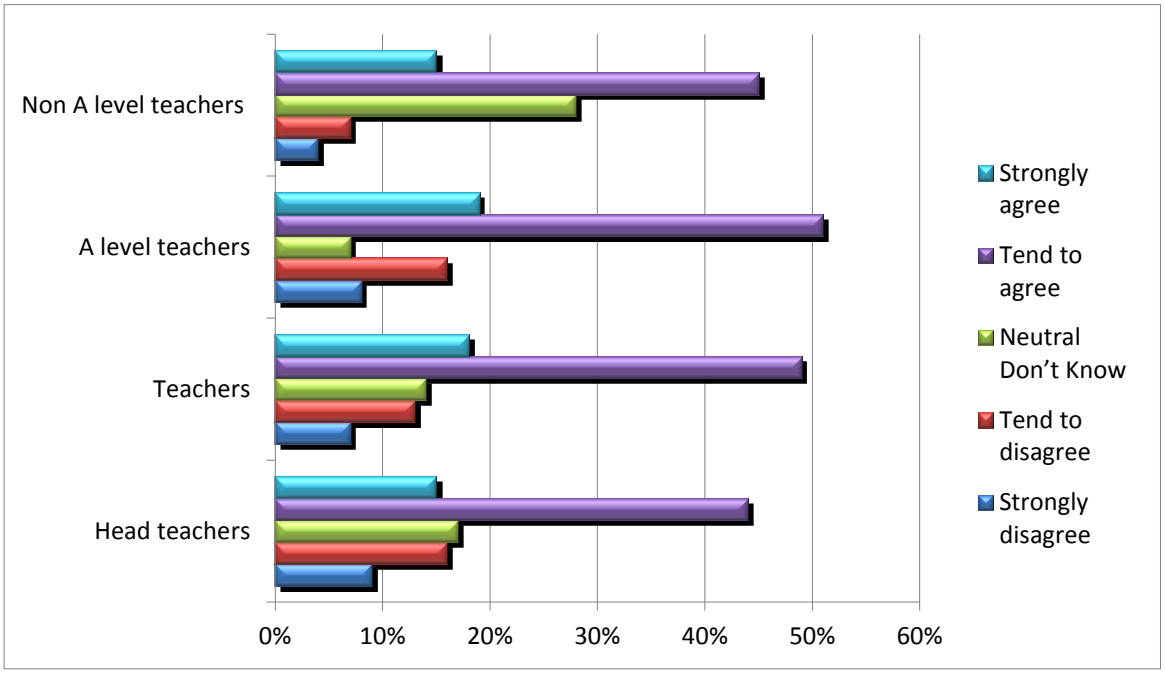
3.2.1 Public perceptions of marking

One useful source of information on public perceptions of marking comes from our recent reliability programme. In 2010, a series of several focus groups and surveys was completed with teachers, employers, parents, students and the wider public, generating rich and robust information. This consultation found that the public has significant trust that the exam system awards candidates the outcomes that they deserve. The public believes that examiners are subject experts and that they award marks fairly. However, given the scale of the system, there is recognition that some degree of human error is inevitable. The public also recognises that the system rests on expert judgement and that some subjects require more interpretation and subjectivity than others (Chamberlain, 2010b; He et al., 2010).

We also carry out an annual perceptions survey to gather stakeholder opinion on issues relating to qualifications in England. Over the years, these perceptions surveys have shown that marking is an important factor in teachers' confidence in qualifications. Amongst head teachers, marking was the most frequently mentioned concern about GCSE qualifications in 2012.

In 2012 we found that the majority of teachers and head teachers were confident in the quality of marking of A levels (and, to a lesser extent, GCSEs). However, there was a sizeable minority who did not share this confidence. One in five teachers (20 per cent) and a quarter of head teachers (25 per cent) were not confident in the accuracy of A level marking.

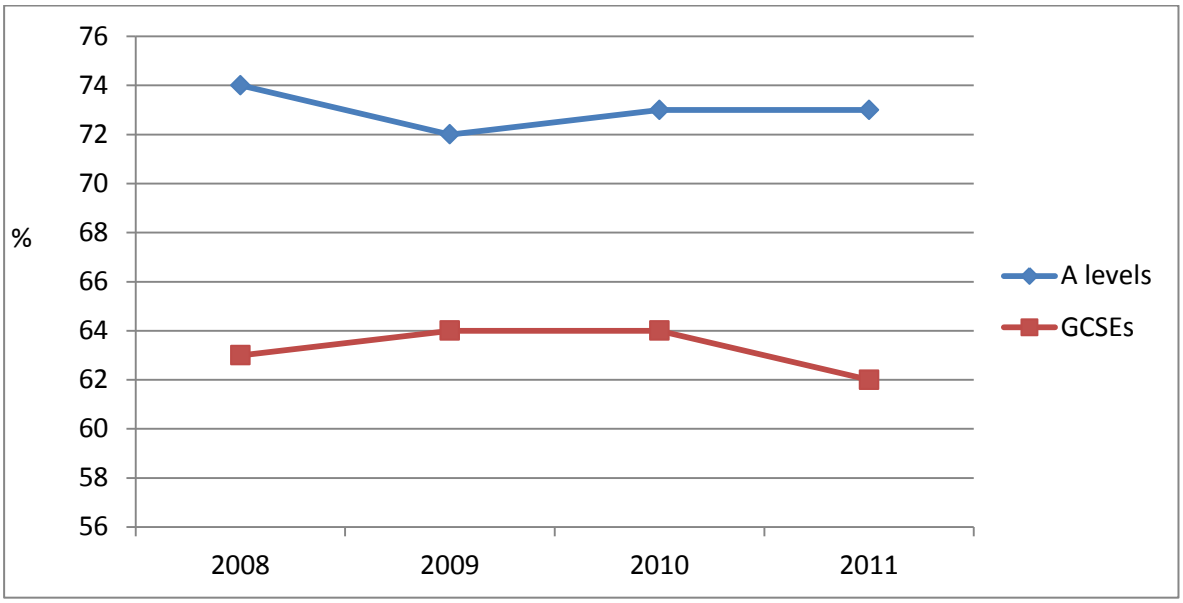
To what extent do you agree or disagree with the following statement: "I have confidence in the accuracy of the marking of A level papers"



Effective base: 170 head teachers and 498 teachers, including 332 who teach A levels and 169 who do not teach A levels, in England, by telephone, for Ofqual (Nov to Dec 2012).

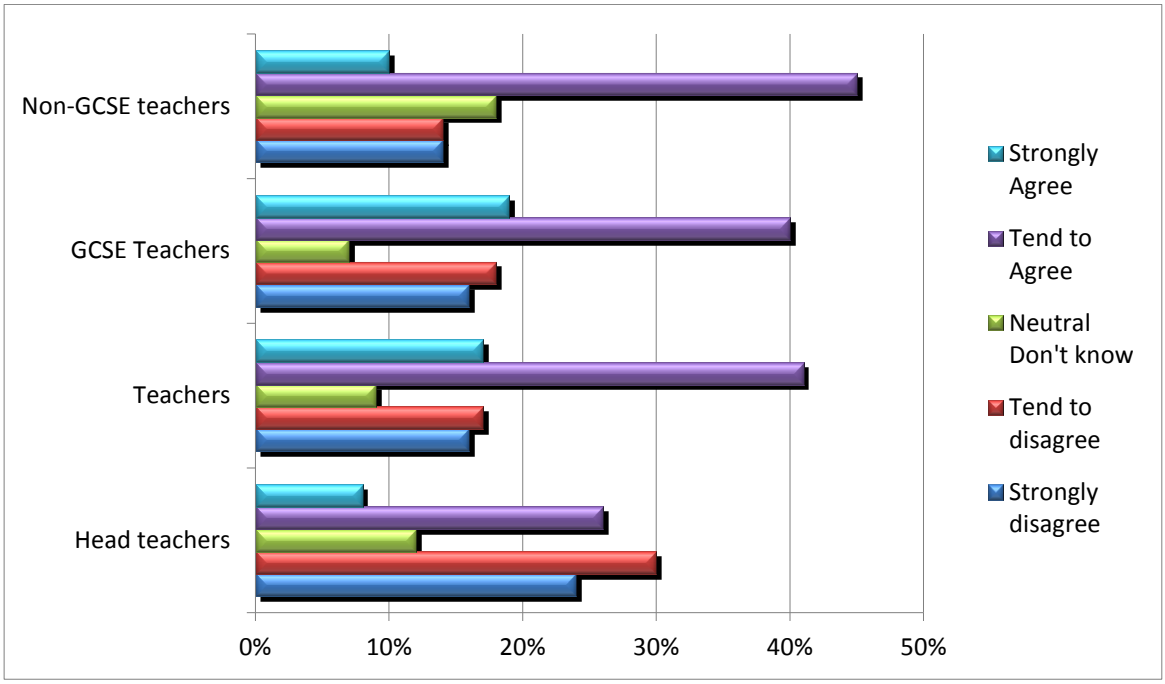
We cannot meaningfully compare the perceptions of teachers in 2012 to those of previous years due to significant changes to the methodology used in the perceptions survey. The graph below shows that up until the changes to the survey in 2012, confidence in the marking of A levels had been stable, with between 72 per cent and 74 per cent of teachers confident in A level marking. However, confidence in GCSEs dropped slightly from 64 per cent to 62 per cent between 2010 and 2011.

"I have confidence in the accuracy and quality of the marking of A level papers/GCSE papers" (Tend to agree and Strongly agree)



In 2012, as for previous years, confidence in the GCSE system was lower than for the A level system. The main causes of concern were varied, and they reflect the impact that the dissatisfaction around the grading of GCSE English in 2012 has had on perceptions of the system. In particular, head teachers were significantly less confident than GCSE teachers in the accuracy of GCSE marking. Just a third of head teachers were confident about the issue (34 per cent) compared with 59 per cent of GCSE teachers. Over half of head teachers (54 per cent) and a third of teachers (34 per cent) were not confident in GCSE marking.

To what extent do you agree or disagree with the following statement: "I have confidence in the accuracy of the marking of GCSE papers"



Effective base: 170 head teachers and 498 teachers, including 423 who teach GCSEs and 79 who do not teach GCSEs, in England , by telephone, for Ofqual (Nov to Dec 2012).

Head teachers were also considerably more likely than teachers to feel that the accuracy of GCSE marking had declined over the past year (64 per cent of head teachers compared with 40 per cent of GCSE teachers felt this to be the case).

In 2012, students had more confidence than parents in A level marking. Thirteen per cent of students and 15 per cent of parents were not confident in A level marking. This represents a slight increase from the 2011 figures, where 9 per cent of both students and parents were not confident in A level marking. In contrast, 25 per cent of parents and 15 per cent of students were not confident in GCSE marking. This is up from 9 per cent and 8 per cent respectively in 2011.

3.2.2 Stakeholder concerns about quality of marking

Over recent months, we have been contacted by some organisations to inform us of their own concerns about quality of marking in GCSEs and A levels. We welcome these contributions and are acting upon them.

In September 2012, the Headmasters' and Headmistresses' Conference (HMC) published a report listing a number of perceived failings in the marking of exams. It states that numerous HMC schools have experienced inconsistent marking in certain GCSEs, level 1 and 2 certificates (known as IGCSEs) and A levels, and seen

unexpected patterns in marks across whole classes. As a result, they have challenged exam boards about the performance of “rogue” markers. In some cases this has resulted in candidates receiving new marks that are significantly higher than the original mark. HMC suggests that poor examiners and the inconsistent application of mark schemes may be behind some of these marking issues.

In its report, HMC makes a series of suggestions for us on quality of marking. We welcome this input and have incorporated it into the scope of this project. For example, we are scrutinising in detail the marking processes and checks in place at exam boards, as well as the people involved in the process.

3.2.3 Enquiries about results

We are also aware that some stakeholders lack confidence in the EAR process by which schools challenge exam boards if they believe a mark is incorrect. Once they receive an EAR, exam boards review the marking to ensure that the mark scheme was applied correctly and that no clerical errors were made. We are aware of some concerns about the timeliness, cost and transparency of this process. HMC also provides some specific criticisms of the way in which EARs are processed by exam boards. It suggests that EARs should comprise a full re-mark of candidate scripts, rather than a review of the marking. We will carry out a separate review into the EAR and appeals processes to report in August 2013.

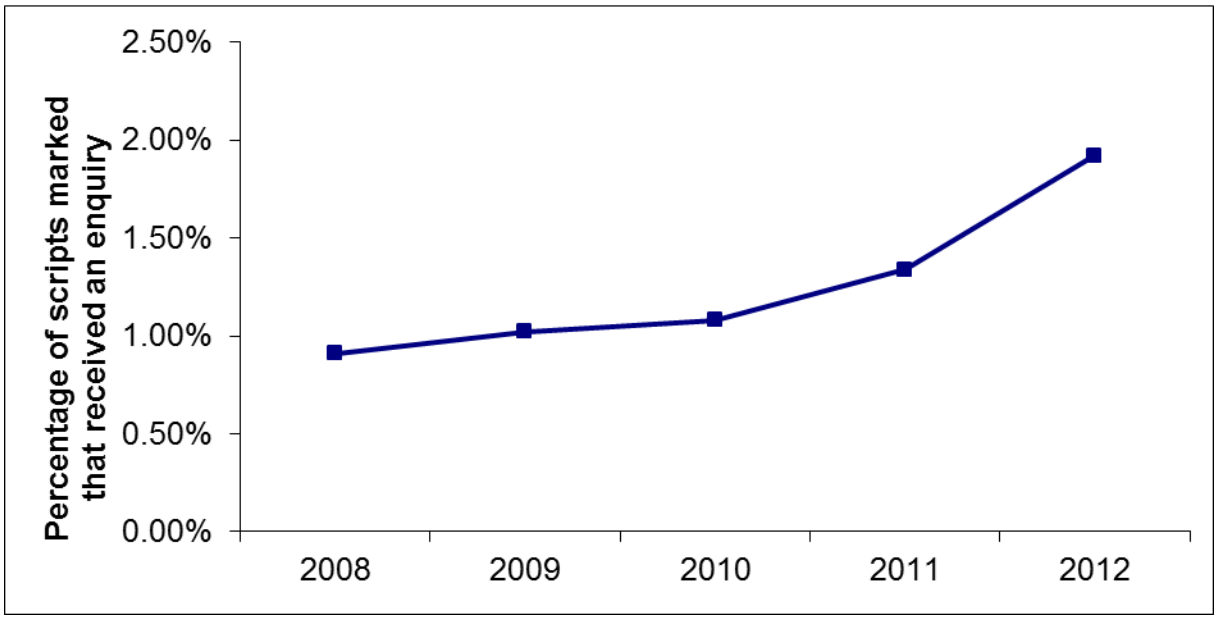
Data on the number of EARs gives an indication of how the marking system is performing more widely, or, at the very least, a picture of how schools perceive the marking system. However, numbers of EARs cannot necessarily be used as a definitive indicator of marking quality because of the way they are used by schools and colleges. Evidence from exam boards shows that EARs are not used consistently by schools – they are often routinely submitted for candidates scoring a mark just below certain key grade boundaries, for example. They are also used far more heavily by some schools and colleges than others.

We publish an annual statistical report on EARs.[19] Between 2011 and 2012, the number of EARs submitted increased by 36 per cent (compared with an increase in exam entries of 15 per cent). This figure was unusually high and likely to have been driven to a degree by the dissatisfaction around the grading of GCSE English. In 2012, a total of 275,808 marking checks[20] were requested for GCSE and A level exams, accounting for 1.9 per cent of all exam scripts. This percentage has

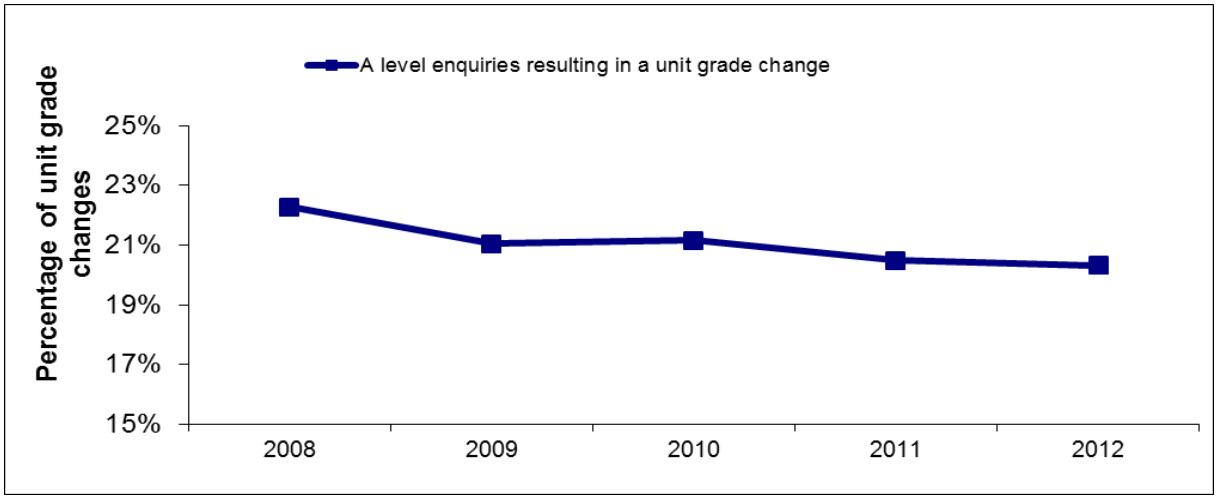
[19] www.ofqual.gov.uk/standards/statistics/enquiries-about-results

[20] Service 1 re-check or a service 2 re-mark.

increased from 1.34 per cent in 2011 and has shown a steady increase over the last five years. In 2008 it was 0.91 per cent.



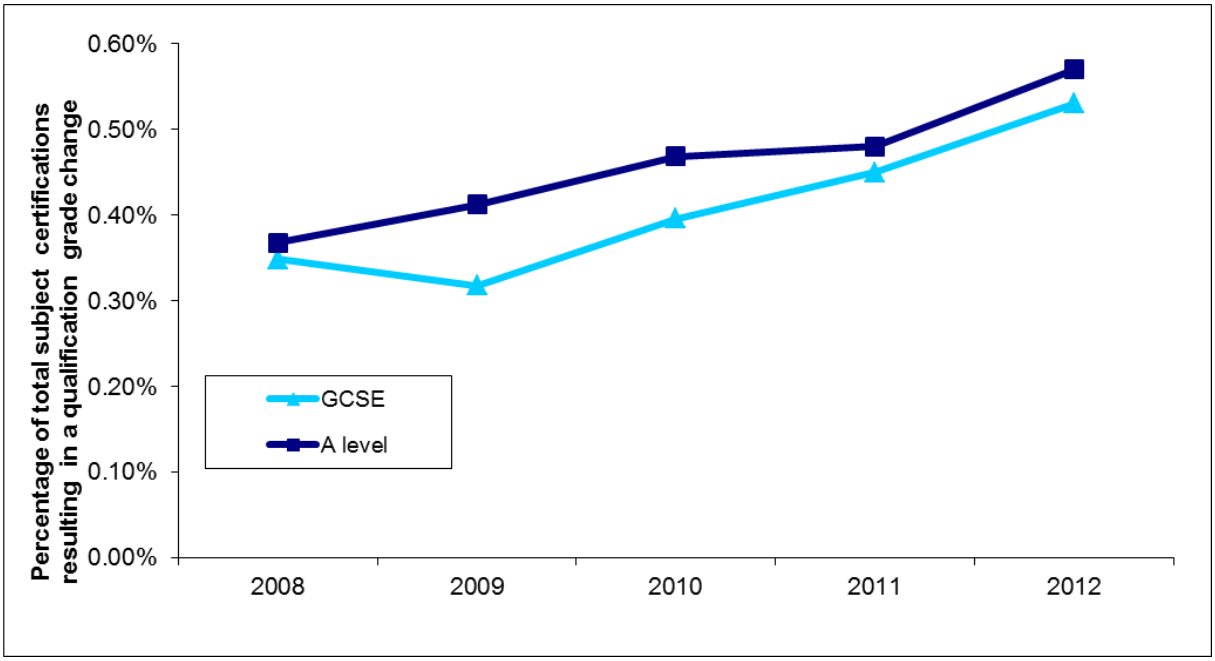
Of these EARs, the majority did not change the outcome for the candidate. In summer 2012, some 16 per cent of EARs resulted in changes to a candidate's qualification grade at A level or GCSE. Whilst the total number of EARs has been rising, the proportion of EARs resulting in a grade change has been decreasing over the last five years. The graph below shows the proportion of unit grade changes at A level.



This illustrates the difficulty in relying on EARs as a measure of marking quality – an increase in the number of EARs does not necessarily mean more marking errors.

As a result of the increase in the number of EARs over recent years, the number of qualification grade changes has risen. In summer 2012, 45,600 EARs resulted in a grade change. This represents 0.5 per cent of all subject certifications in GCSEs or A

levels for that exam series. The graph below shows that there has been a steady increase in this percentage at both GCSE and A level. However, the rate of this increase has been slower than the increase in the number of EARs made by schools and colleges.



Whilst we must treat any EAR data with caution, the increase in the number of grade changes is important. We are looking closely at what this tells us about marking quality as well as about the EAR process itself. We will also carry out further analysis to see if we can identify any patterns with perceived marking issues in certain exam boards or across types of subject. It is planned that the results of this analysis will be used to influence and drive improvements in specific units, and subjects, and across exam boards. This work will be combined with our review of the appeal process, including the exam process review service.

4. What does research tell us about improving quality of marking?

In this review, we are taking stock of existing evidence and expert opinion on marking quality. If the factors that affect quality of marking are better understood, this could help improve the reliability and accuracy of marking. Considerable research has been carried out into this topic over the years, and we commissioned the National Foundation for Educational Research (NFER) to complete a literature review to summarise the most recent research on marking quality, and, more specifically, marking reliability.

This section provides a brief discussion of some of the key findings. It explains the factors that affect marking reliability and discusses advances in improving marking reliability. The full report, *A Review of Literature Marking Reliability Research*, is published alongside this report[21].

4.1 Factors influencing marking quality

The literature review shows that a number of factors affect marking quality. Reliability of marking is mainly affected by the style of individual exam questions, the characteristics of candidate responses and the mark schemes used. It is also affected by the characteristics of the examiners involved and the processes that are used for marking.

The single biggest factor influencing the reliability of marking is the type of question used in an exam (Meadows and Billington, 2005e; Tisi, 2013f). As discussed in section 3, in assessment there is always a balance between reliability and validity. Whilst tightly constrained questions result in a higher reliability of marking, in some subjects the use of high-mark questions with extended responses is an important aspect of validity. However, even for those questions that traditionally have lower levels of marking reliability, it may be possible to make improvements.

Changes to mark schemes are one possible means of improving marking quality. Research suggests that small improvements to mark schemes could optimise marking reliability, particularly with the levels-based mark schemes used for less constrained questions. Detailed mark schemes can help examiners to decide more precisely what makes a creditworthy response, and so are important in improving marking consistency. Such mark schemes should give clear principles to help markers discriminate better between good and poorer responses as well as providing detailed guidance and evaluation criteria (Tisi et al., 2013g; Meadows and Billington,

[21] www.ofqual.gov.uk/files/2013-06-nfer-a-review-of-literature-on-marking-reliability-research.pdf

2005f). There are also changes that could be made to the structure of mark schemes in order to improve marking accuracy. This includes designing mark schemes so that mark bands are of equal width (Pinot de Moira, 2012a) and using a greater number of narrower mark bands (Tisi et al., 2013h). These features are already present in the best mark schemes, but we believe that they could be adopted more widely.

In addition to the design of the exam, research shows that the characteristics of markers can affect the marking quality. As we may expect, examiner experience is important, with a number of studies finding that inexperienced examiners marked less accurately than experienced markers. However, for simple, highly constrained questions, markers with no subject or examining experience can mark just as reliably as very experienced examiners (Tisi et al., 2013i).

The marking processes used by exam boards also affect quality of marking. Examiners must be trained to apply a mark scheme consistently to the papers or items that they are marking, and there must be effective quality controls in place to ensure that they mark reliably and accurately throughout the marking period. These processes must also address possible sources of simple clerical errors in marking, such as markers incorrectly adding up or recording candidate marks.

4.2 Recent advances in marking reliability

In recent years, our understanding of the factors influencing marking reliability has improved. Over the same period, research suggests that there have been advances in improving marking reliability in practice (Tisi et al., 2013j). The most significant recent change to the system has been the introduction of on-screen marking. This has enabled changes to marking processes, to the people involved in marking and to the monitoring of these. All these changes have the potential to improve the quality of marking of general qualifications.

As discussed in section 2, on-screen marking also supports item-level marking whereby examiners are assigned a set of items (or questions) to mark, rather than a number of whole exam papers. Existing research shows that item-level marking could improve marking reliability by reducing the effects of biases caused by the rest of the question paper. When one examiner marks all the questions on an exam script, the mark that they allocate to one item may be affected by the candidate's responses to another, unrelated question. In other words, an examiner might carry forward preconceived ideas about the level of a candidate's understanding (whether positive or negative assumptions) based on answers to previous unrelated items. This is known as the "halo" effect and is eliminated in item-level marking (Spear, 1996b).

Item-level marking could also improve reliability to a certain extent by reducing the influence of a single examiner on an exam script. When different examiners mark each item on a script, variations in their marking are likely to cancel each other out.

That is, for each question that is over-marked, there is likely to be one that is under-marked. The more examiners who contribute to the final mark, the more reliable the marking of the overall question paper will be (Pinot de Moira, 2011b). However, the research is not conclusive. A recent research study carried out by Black and Curcin (in prep) compared whole-script and item-level marking on one unit. The study found that whilst whole-script marking can produce a "halo" effect, marking a unit at item level did not have any substantial impact on marking reliability. In this study, however, it did eradicate the (rare) extreme results. We would like to see more research evidence on the pros and cons of item-level marking in practice.

Item-level marking also allows the distribution of exam questions across examiners to ensure that the questions are marked by those with the most appropriate level of expertise. The items that are harder to mark can be sent to examiners with the most expertise (who, in general, will mark complex items more reliably than less experienced examiners), whilst items that are easier to mark can be sent to examiners with the least expertise. Similarly, where optional routes and questions are possible, responses can be sent to examiners with specific subject knowledge. For example, a response on a particular English text can be sent to an examiner who is more familiar with that text.

This distribution of items to the most appropriate examiners is a clear benefit of item-level marking. However, there is little evidence that this takes place routinely in the current system. Where item-level marking is used by exam boards, it is a fairly new innovation in their marking practice and they have not yet exploited it fully. There is still work that can be done to realise the full benefits of this type of marking.

There are other benefits of on-screen marking beyond item-level marking. As discussed, it is likely to improve marking quality by allowing more regular marker monitoring by exam boards. Rather than sampling work at two or three points throughout the marking period, on-screen marking can be done in real time to allow continuous support and feedback to examiners. Any inconsistent or inaccurate marking can be detected earlier, allowing the examiner to be retrained or stopped from marking.

All the advantages of on-screen marking could be seen as irrelevant if something about the process made it inherently less reliable than paper-based marking. However, the limited empirical evidence that does exist suggests that on-screen marking is as reliable as traditional forms of marking. Some studies suggest that the cognitive workload of examiners is higher when trying to read and mark work on-screen. However, a small empirical research study in 2012 showed that this does not affect the accuracy or reliability of their marking, either for short answer or essay questions (Johnson et al., 2012).

On-screen marking also lends itself more easily to double or multiple marking. In double marking, two examiners independently assess each script (or item). The final mark is the combination of two separate marks. In multiple marking, more than two examiners are used. The combination of double/multiple marks to produce a final score is an acknowledgement that legitimate differences in opinion can exist between examiners. This is fundamentally different from the current hierarchical system, in which the marks of the most senior examiner are considered to be the most “true”.

As early as the 1940s, studies of double and multiple marking suggested that they could increase the reliability of marking (Meadows and Billington, 2005g). Further empirical and theoretical studies in the 1960s and 1970s supported the claims that multiple marking is more reliable than a single mark (Tisi, 2013k). This led to the use of double marking by a number of exam boards in the 1960s and 1970s in some of the more subjective assessments.

Recent research still indicates that there are gains to be made to reliability through double marking. This shows that the greatest increases in reliability come from increasing the marking team from one to two (Kim and Wilson, 2009). A bigger team increases marking reliability further, although the gains here are much smaller. It is also best targeted at more subjective question types. For a full discussion of this research, see Tisi et al (2013l) or Meadows and Billington (2005h).

The use of double marking by exam boards has diminished since the 1960s and 1970s. For all its benefits, there are significant logistical and financial obstacles to using it more widely. Not least, recruiting the number of examiners that would be required for double marking may be prohibitive. It is by no means certain that the benefits of double marking would outweigh the considerable costs (Fearnley, 2006).

This said, there are still some small-scale examples of double marking being used in the marking of academic qualifications in England. As discussed in section 3, AQA and WJEC both use double marking of a sample of items in more subjective disciplines as a means of monitoring marking reliability.

4.3 What does this tell us?

The literature review shows that there has been progress in marking practice. However, we believe that there are still opportunities to improve quality of marking further. Research shows that the biggest gains in marking reliability would come from modifying features of questions used in exams (Meadows and Billington, 2005i). As we have already discussed, some of these changes would lower the validity of the assessment and would be incompatible with current reforms to GCSEs and A levels. However, even within these constraints, research suggests that it may be possible to make further improvements by increasingly:

- ensuring that mark schemes are sufficiently detailed to help examiners decide precisely what makes a creditworthy response; such mark schemes should give multiple examples of the sorts of responses that the question might provoke and, crucially, provide examiners with principles for discriminating between different levels of responses;
- selecting markers with particular characteristics for different item types, for example using the most experienced examiners on items that are complex to mark;
- using item-level marking so that more than one marker contributes to a candidate's overall mark rather than a single marker assessing a whole script; this reduces the effects of random error and removes the "halo" effect;
- using on-screen marking, which allows continuous marker monitoring that enables inaccuracies to be detected early and corrected; on-screen marking should also reduce errors resulting from the incorrect addition or transcription of marks;
- using blind re-marking as part of the monitoring of examiners, as re-marking with the comments and marks visible on the script is likely to underestimate unreliability;
- using multiple marking to improve marking reliability for some question types.

(Tisi et al., 2013m)

We know that some of these changes (such as on-screen and item-level marking) are already being implemented to an extent by exam boards. Others present opportunities that can be considered in the future.

5. Next steps

This report is the first stage in our review programme of the quality of marking in exams in A levels, GCSEs and other academic qualifications. The programme is in three parts:

1. to build up a picture of the current marking system;

2. to review the EAR and appeals processes;

3. to identify where improvements can be made to the quality of marking in exams in A levels, GCSEs and other academic qualifications.

The second stage of this work is a review of the EAR and appeals processes used by schools to challenge the marks given to candidates. We will report on this in early August 2013.

The final report, due in the autumn, will take a holistic view of the system to identify where any improvements could be made to the marking of exams in England. Over the coming months we will continue to gather detailed evidence to support this work. To understand marking quality, we believe that it is necessary to focus on six themes that are central to the marking of exams. These are: the marking process; the people involved; metrics (how quality of marking is evaluated by exam boards during marking); stakeholders' perceptions and expectations of marking; and mark scheme design and any constraints that the exam boards are working within. For a more detailed list of research questions, see appendix E. As we set out in this report, the biggest factor influencing reliability of marking is the design of the assessment itself. We will focus on this elsewhere – in our work to support qualification reforms.

Our initial review of the evidence has identified a number of areas that we are particularly interested in exploring in more detail. These include:

- the benefits and drawbacks associated with double marking;

Research indicates that there are gains to be made to reliability through double marking in more subjective disciplines and question types. However, for all its benefits, there are significant logistical and financial obstacles to the use of double marking. We now need to consider whether double marking is feasible and whether its benefits outweigh the drawbacks involved.

- the benefits and drawbacks of item-level marking;

Research suggests that, at least in theory, item-level marking has the potential to improve the accuracy of marking in exams by reducing the effects of biases caused by the rest of the question paper (the "halo" effect) and by removing the influence of

a single examiner on an exam script. However, we believe that more evidence needs to be gathered to show the benefits and drawbacks of item-level marking in practice.

- the impact of different methods of standardisation on marking quality;

The shift from face-to-face to online standardisation has been one of the more controversial developments in marking practice in recent years, and more work needs to be done to assess the impact that this has had on both marking quality and in the engagement of examiners in marking.

- the use of common tolerances across exam boards;

The use of tolerances recognises that there can be legitimate differences in professional judgement between examiners, particularly in certain subjects or question types. In a qualifications market we can accept variation and innovation in practice. However, there are instances in which consistency across exam boards is desirable. We will consider the case for the use of common tolerances across exam boards to monitor the work of examiners.

- establishing a common suite of metrics to measure quality of marking across exam boards;

At present there is no single accepted way of measuring marking quality. Very few common metrics are available that allow us to measure quality in the exam system in England. We will work with exam boards to investigate the feasibility of establishing some common metrics of marking quality.

- the role of mark schemes in ensuring marking reliability;

Research suggests that small improvements to mark schemes could optimise marking reliability. Detailed mark schemes can help examiners to decide more precisely what makes a creditworthy response, particularly if they provide clear principles to help markers discriminate better from poorer responses. Over the coming months, we will further investigate the role of enhancements to mark schemes in improving marking quality.

- stakeholder perceptions of marking quality;

Whilst our consultation indicates that confidence in the marking of GCSEs and A levels is generally high, we recognise that a significant minority of teachers do have concerns about the marking of exams. To explore these concerns further, we are consulting teachers and head teachers via an online survey on our website[22] as well

[22] http://ofqual.gov.uk/news/we-want-teachers-views-on-quality-of-marking

as through in-depth interviews and focus groups. Analysis of our consultation findings will be presented in our final report. Alongside this, we will also present the results from our survey and focus groups with examiners.

- any issues with marking quality in specific subjects or exam boards.

We will continue to gather data to analyse whether there are any issues with marking reliability in specific subjects or exam boards. Additionally, we know that exam boards rely on tightly controlled marking processes and quality controls to manage the volumes of exam scripts within the deadlines involved. We are studying all aspects of these arrangements to assess where they work well, where they don't work so well and where there might be room for improvement.

5.1 Methodology

We are gathering detailed evidence to answer our original research questions and address the emerging themes above. Much of this evidence is dependent on gaining information from the seven exam boards that provide these qualifications. As well as these exam boards, we are consulting more widely with examiners, national organisations and government bodies.

For more detail on our evidence-gathering activities see appendix E.

6. References

Black, B. and Curcin, M. (in prep) *Marking Item by Item Versus Whole Script – What Difference Does It Make?* Cambridge Assessment internal report.

Bramley, T. (2008) *Mark Scheme Features Associated with Different Levels of Marker Agreement*. Cambridge Assessment. Available at: www.cambridgeassessment.org.uk/ca/digitalAssets/171183_TB_MSfeatures_BERA08.pdf. (accessed 14th May 2013).

Brooks, V. (2004) *Double Marking Revisited*, *British Journal of Educational Studies*, volume 52, issue 1, pages 29 to 46.

Chamberlain, S. and Taylor, R. (2010) *Online or Face-to-face? An Experimental Study of Examiner Training, British Journal of Educational Technology*, volume 42, issue 4, pages 665 to 675, July 2011.

Chamberlain, S. (2010) *Public Perceptions of Reliability*. Opposs, D. and He, Q. (eds.) *Ofqual's Reliability Compendium*, pages 691 to 725. Coventry, Ofqual.

Crisp, V. (2010). *Towards a Model of the Judgement Processes Involved in Examination Marking*, *Oxford Review of Education*, volume 36, issue 1, pages 1 to 21.

Fearnley, A. and Billington, L. (2006) *An Investigation of Targeted Double Marking for GCSE and GCE*. AQA internal report.

Gove, M. (2013) Ofqual policy steer letter: *Reforming Key Stage 4 Qualifications*. Available at: www.ofqual.gov.uk/files/2013-02-07-letter-from-michael-gove-reform-of-ks4-qualifications.pdf (accessed 17th April 2013).

Hayward, G., Sturdy, S. and James, S. (2005) *Estimating the Reliability of Predicted Grades.* University and Colleges Admission Service (UCAS). Available at: www.ucasresearch.com/documents/Predicted_Grades_2005.pdf (accessed 21st May 2013).

He, Q., Opposs, D. and Boyle, A. (2010) *A Quantitative Investigation into Public Perceptions of Reliability in Examination Results in England*. Opposs, D. and He, Q. (eds.) *Ofqual's Reliability Compendium*, pages 727 to 767. Coventry, Ofqual.

HMC (2012) *England's "Examinations Industry": Deterioration and Decay*. Available at: www.hmc.org.uk/wp-content/uploads/2012/09/HMC-Report-on-English-Exams-9-12-v-13.pdf (accessed 17th April 2013).

IBO (2012) *What Are the Dates for the 2013 DP Examinations?* Available at: https://ibanswers.ibo.org/app/answers/detail/a_id/3176/kw/exam%20timetable%202013/session/L3RpbWUvMTM2NTc1NzcwNy9zaWQvYlVVM2d3bmw%3D (accessed 17th April 2013).

Ipsos MORI (2013) *Perceptions of A levels, GCSEs and Other Qualifications: Wave 11*. Available at: www.ofqual.gov.uk/files/2013-05-03-perceptions-A-levels-GCSEs-and-other-qualifications–wave11-perceptions-teachers-general-public.pdf (accessed 14th May 2013).

JCQ (2012) *Key Dates in the Examination Cycle, 2012/2013*. Available at: www.jcq.org.uk/exams-office/key-dates-and-timetables (accessed 17th April 2013).

Johnson, M., Hopkin, R., Shiell, H. and Bell, J.F. (2012) *Extended Essay Marking on Screen: Is Examiner Marking Accuracy Influenced by Marking Mode? Educational Research and Evaluation*, volume 18, issue 2, pages 107 to 124.

Kim, S.C. and Wilson, M. (2009) *A Comparative Analysis of the Ratings in Performance Assessment Using Generalizability Theory and the Many-facet Rasch Model*, *Journal of Applied Measurement*, volume 10, issue 4, pages 408 to 423.

Meadows, M. and Billington, L. (2005) *A Review of the Literature on Marking Reliability, Report to NAA.* National Assessment Agency. Available at: https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf (accessed 17th April 2013).

Meadows, M. and Billington, L. (2007) *NAA Enhancing the Quality of Marking Project: Final Report for Research on Marker Selection*. London, QCA. Available at: http://archive.teachfind.com/qcda/orderline.qcda.gov.uk/gempdf/184962531X/QCDA104980_marker_selection.pdf (accessed 20th May 2013).

Nunnally, J. and Bernstein, I. (1994) *Psychometric Theory*. New York, McGraw-Hill.

Ofqual (2011a) *Inquiry into Examination Errors Summer 2011 Final Report*. Available at: http://ofqual.gov.uk/files/2011-12-20-inquiry-into-examination-errors-summer-2011-final-report.pdf (accessed 20th May 2013).

Ofqual (2011b) *Reliability*. Available at: www.ofqual.gov.uk/standards/reliability (accessed 17th April 2013).

Ofqual (2012a) Ofqual issues Direction to exam board following clerical errors in 2011. Available at: www.ofqual.gov.uk/news-and-announcements/130-news-and-announcements-press-releases/977-ofqual-issues-direction-to-exam-board-following-clerical-errors-in-2011 (accessed 17th April 2013).

Ofqual (2012b) *International Comparisons in Senior Secondary Assessment Full Report*. Available at: www.ofqual.gov.uk/standards/research-reports/92-articles/23-comparability#international (accessed 24th May 2013).

Ofqual (2012c) *Statistical Bulletin for Enquiries about Results for GCSE and A level: Summer 2012 Exam Series*. Available at: www.ofqual.gov.uk/files/2012-12-06-statistical-bulletin-enquiries-about-results-for-gcse-and-a-level.pdf (accessed 17th April 2013).

Pinot de Moira, A. (2011a) *Levels-based Mark Schemes and Marking Bias*. Centre for Education Research and Policy. Manchester.

Pinot de Moira, A. (2011b) *Why Item Mark? The Advantages and Disadvantages of E-marking*. Manchester, AQA, Centre for Education Research and Policy.

Spear, M. (1996) *The Influence of Halo Effects upon Teachers' Assessments of Written Work*. *Research in Education*, page 85.

Tisi, J., Whitehouse, G., Maughan, S. and Burdett, N. (2013) *A Review of Literature on Marking Reliability Research* (report for Ofqual). Slough, NFER.

Appendix A – Internal assessment

Assessment profiles for GCSEs as specified in our GCSE subject criteria:

Subject	External exam	Internal assessment
Maths	100%	0%
English literature History Geography Science	75%	25%
Art and design English language Modern foreign languages	40%	60%

Assessment profiles for A levels as specified in our subject criteria:

Subject	External exam	Internal assessment
Geography Modern foreign languages	100%	0%
Maths	80 – 100%	0 – 20%
History	80 – 85%	15 – 20%
Science*	70 – 80%	20 – 30%
English literature	60 – 85%	15 – 40%
Art and design	0%	100%

*all units in psychology must be externally assessed at both AS and A level

Appendix B – Mark schemes

This appendix provides a brief description of the different types of mark schemes mentioned in this report. It draws heavily on the *Review of Literature on Marking Reliability Research* (Tisi et al., 2013n).

Objective/constrained mark scheme

Items that are objectively marked require very brief responses and greatly constrain how candidates must respond. An unambiguous correct answer exists for the question which that can be completely defined in the mark scheme. The distinction between right and wrong is completely transparent and the marker does not need to use any subjectivity. Examples include multiple-choice questions, answers in the form of a single word or number, and questions that require the indication or identification of information on the question paper (for example, indicating an area on a diagram). Objective mark schemes can be applied with a high degree of accuracy.

For example:

Name the capital city of Finland.

or

Write the chemical symbol for sodium.

Points-based mark schemes

These items usually need responses ranging in length from a few words to one or two paragraphs, or a diagram or graph. Points-based mark schemes list objectively identifiable words, statements or ideas. Marks are awarded for each creditworthy point in the candidate's response. There is generally a one-to-one correspondence between the number of correct answers that the candidate gives and the number of marks that should be awarded. All the creditworthy points are listed in the mark scheme, but the marker still needs to find the relevant elements in the response.

One criticism of this type of mark scheme is that the relative importance of different statements is rarely addressed – every point is treated as equal in value. Therefore, if the maximum mark is lower than the number of creditworthy points, a candidate can achieve full marks even if they omit key parts of the answer. Similarly, the tactic of simply writing down everything that comes to mind, even if it is not relevant, can achieve high marks without candidates fully understanding what they are writing.

Levels-based mark schemes

These items usually require longer answers, ranging from one or two paragraphs to multiple-page essays. Levels-based mark schemes divide the mark range into

several bands, each representing a distinguishable level of quality of response. The level descriptors may include features of language, content or both.

In a holistic levels-based scheme, markers make an overall judgment of the performance. Each level may include a number of different response features, but no explicit weightings are given to the different features. Therefore, if a response merits different levels for different aspects, the marker must use their judgment to decide the best-fit category, without explicit information about which aspects are most highly valued. The result is that different markers may award different marks because they have different understandings of what it means for the response to be good. Alternatively, markers may award the same mark for different reasons. These issues both undermine the construct-validity of the test, that is, the same marks may not mean the same thing in terms of the trait that the test is supposed to measure.

Analytic levels-based mark schemes separate the aspects of interest and provide level descriptors, and associated mark bands, for each aspect. That is, they explicitly weight the different features of response.

Holistic scoring is rapid but only a single score is reported, thus the same score assigned to two separate pieces of work may represent two entirely distinct sets of characteristics. In contrast, in analytic scoring a mark is awarded to each of a number of different aspects of the task. It is, therefore, much slower than holistic marking but provides more information about the candidate's ability.

Comparative studies of the reliability of the different marking methods show that analytic marking is more reliable than holistic marking in some cases, but that there is no difference in other cases. Analytic marking is more labour intensive, so, in terms of time and cost, several holistic markers are equivalent to one analytic marker, and there is some evidence that the pooled results of a set of holistic markers are more reliable than the results of one analytic marker (Tisi et al., 2013o).

Appendix C- Journey of a script

Marking process – paper marking

1. Candidate sits paper exam.
2. The script is couriered to an examiner or senior team.
3. A sample of scripts is marked by the senior examining team. These scripts will be used for training examiners at the standardisation stage.
4. Standardisation – examiners meet to work through practice scripts to ensure they are fully competent in applying the mark scheme to the correct standard before they begin marking.
5. Qualifying/approval – examiners mark a first sample of scripts. These are reviewed by their team leader. If they are marking to the required standard they are approved to begin live marking.
6. Examiners begin marking their batch of exam scripts.
7. Monitoring – during live marking, markers send one or more samples of marked scripts to their team leader. If any serious issues are found with their marking they may be stopped from marking at this point.
8. At the end of marking, examiners send all the marked scripts to the exam board.
9. Clerical checks are performed on scripts by the exam board to confirm that the marks are correctly totalled and recorded.
10. Marking review – some scripts will be reviewed again if there are any concerns about an examiner. They may be re-marked or have a statistical adjustment (scaling) applied.
11. Marking is completed.
12. The scripts are securely stored for 6 to 12 months before being destroyed.

Marking process – e-marking

1. Candidate sits paper exam.
2. The script is couriered to a UK scanning centre within an agreed time scale.

3. The scripts are checked and prepared for scanning. The hardcopy originals are stored.

4. Scans are checked for completeness.

5. As scripts are scanned and checked, they are released into the marking system.

6. The script images are either distributed to examiners as a whole and marked as a complete script or split and marked as individual items (questions).

7. A sample of scripts or items is marked by the senior examining team. These scripts will be used for training examiners at the standardisation stage. They are also used as "seed" items or scripts to monitor examiner performance during marking.

8. Standardisation – examiners work through practice scripts or items to ensure they are fully competent in applying the mark scheme to the correct standard before they begin marking. Standardisation takes place as a meeting or online.

9. Qualifying/approval – examiners mark a first set of scripts or items. These are reviewed by their team leader. If they are marking to the required standard they are approved to begin live marking.

10. Examiners begin marking live exam scripts or items.

11. Images are marked within an agreed time frame. Examiners view a script (or item) on screen and enter the mark given to each question into the online system. Online systems automatically total and check the marks given.

12. Monitoring – the examiner receives "seed" items at regular intervals throughout live marking. These have been pre-marked by senior examiners. If the examiner's mark does not show sufficient agreement with the senior examiners' mark on a number of occasions, there may be additional checks on the examiner, after which he or she receives further guidance or training, or may be stopped from marking.

13. Marking is completed.

14. The scanned scripts can be stored as long as required.

15. The original paper scripts are securely stored for 6 to 12 months before being destroyed.

Appendix D – Uniform mark scale

In modular (unitised) qualifications, candidates sit their units at different times in the course. They will, therefore, sit different papers depending on whether they entered in January or June.

Since the papers are different, it's likely that the level of demand will vary. Awarders will take that into account and set grade boundaries at different raw marks, to reflect the relative level of demand of the papers. The value of the raw marks – the marks the candidate scored on the paper – is dependent on the level of demand of that paper. For example, the same mark on a more demanding paper represents better performance. Therefore, it wouldn't be fair to candidates simply to add up the raw marks to give the overall result, as the candidates taking the more demanding paper would, therefore, be at a slight disadvantage.

The uniform mark scale (UMS) puts all these raw marks on the same scale. The raw mark is converted to a UMS mark that reflects the position of that mark relative to the grade boundaries set in that unit.

For instance, in a unit in January the boundary (minimum) mark for a grade A might be 78, compared with 76 in June. A candidate scoring 77 in January would be just below the number of marks required for a grade A. However, a candidate scoring 77 raw marks in June would achieve a grade A. The UMS mark awarded to these candidates would reflect that difference.

Unit 3	Grade A boundary (raw mark)	Grade A boundary (UMS)
January 2010	78	80
June 2010	76	80

UMS marks from all the units are added together to give an overall mark for the qualification. This conversion ensures the final mark reflects the standard needed to achieve a particular grade.

A detailed explanation of UMS marks and the relationship to grades has been published by AQA in its booklet *Uniform Marks in A level and GCSE Exams and Points in the Diploma*. This is available on its website.[23]

[23] http://store.aqa.org.uk/over/stat_pdf/UNIFORMMARKS-LEAFLET.PDF

Appendix E – Scope and methodology

Aims and scope

The final report will be published in the autumn. The review is investigating the quality of marking in exams in A levels, GCSEs and other academic qualifications in England. There are three aims of the project:

1. improve public understanding of how marking works and its limitations;

2. identify where current arrangements work well (and where they don't);

3. recommend improvements where they might be necessary.

The marking of internal assessment by teachers is not in scope.

What do we mean by quality of marking?

The term "quality of marking" is broad. For the purposes of this review, quality of marking is defined as the accuracy and reliability of marking. This is to say that candidates are given a mark that is as close to their correct, "true" score as possible, and that this is the case no matter who marked their work.

To understand marking quality, we believe that it is necessary to focus on six themes that are central to the marking of exams. These are: the marking process; the people involved; metrics (how quality of marking is evaluated by exam boards during marking); stakeholder perceptions and their expectations of marking; mark scheme design; and any constraints that the exam boards are working within.

Under each theme, we are addressing a number of specific questions, which include:

The marking process:

- What specific marking processes and quality controls are used by exam boards and how do these vary for different subjects and marking methods?

- What impact do the different methods of marking (including item-level marking) have on the quality of marking?

- What impact do the different methods of standardisation have on the quality of marking?

- Where are the risks in the system and what opportunities are there to improve marking processes?

- What are the benefits and costs associated with double marking?

- How desirable are common tolerances across exam boards to monitor the work of examiners?

The people involved:

- Who are the examiners of external exams and what level of subject, teaching and examining experience do they have? What training do they receive?
- How do examiners feel about their role and the level of time commitment and pressure associated with it?
- Are there marker shortages in any areas and, if so, what are the causes of these?
- How are markers evaluated and how is poor performance managed?

Metrics (how quality of marking is evaluated by exam boards during marking):

- How do exam boards measure and evaluate marking (in real time and after the process has been completed)?
- What is the feasibility of establishing a common suite of metrics to measure quality of marking across exam boards?
- Can patterns of EARs give us any insight into quality of marking?

Stakeholder perceptions and their expectations of marking:

- What are the public perceptions of the marking system?
- What levels of understanding of the marking of external exams do teachers have?

Mark scheme design:

- What is the role of mark schemes in ensuring marking reliability?

Constraints:

- What constraints on quality of marking are imposed by the wider exam or education system?
- What resource constraints on quality of marking are there within exam boards?

We also know that the biggest factor influencing reliability of marking is the design of an exam: the style of the questions and mark schemes. Given this, our final report will also consider the role of mark schemes in ensuring marking reliability.

Methodology

We are gathering detailed evidence to answer the questions above. Much of it is dependent on gaining information from the seven exam boards that provide these qualifications. As well as these exam boards, we are consulting more widely with examiners, teachers and head teachers, and national organisations and government bodies. The evidence-gathering activities are detailed below.

Literature review

We commissioned NFER to produce a literature review on marking reliability. This synthesises the latest research on best practice in marking reliability and summarises any recent advances in marking reliability. Findings from this literature review have been discussed throughout this report.

Exam board visits

We have completed our first round of visits to exam boards and collected detailed evidence on marking processes and the people involved. Further visits will provide information on the metrics that exam boards use to monitor and evaluate quality of marking, as well as the constraints that they work within. We will analyse the data to investigate differences in approaches between exam boards.

Statistical analysis – internal and external

One of our first activities was to carry out a stocktake of our existing data, such as EAR figures, complaints and the results of previous consultations. Further exploration of these datasets will be carried out over the coming months. In particular, EAR data will be interrogated to see how EARs vary by subject, exam board, qualification type and the type of marking used.

Data is also being collected from exam boards to show how the system works in practice and how quality of marking is measured by each board. Where possible, data will be analysed to investigate differences in approaches between (and within) boards, as well as how marking approaches have changed over time.

Stakeholder consultation

Whilst our consultation indicates that confidence in the marking of GCSEs and A levels is generally high, we recognise that a significant minority of teachers do have concerns about the marking of exams. To explore these concerns further, we are consulting teachers and head teachers via an online survey on our website[24] as well as through in-depth interviews and focus groups.

[24] http://ofqual.gov.uk/news/we-want-teachers-views-on-quality-of-marking

In May 2013 we carried out a survey with examiners that mark GCSEs, A levels and other academic qualifications. This has helped us to understand who the examiners are and what experience they have. It will also gather their perceptions of the marking process and any improvements that they would like to see made to the system. Some very early findings from this survey have been presented in section 2

We will also approach national organisations to capture their insights about the marking system, where they believe that current arrangements work well and where they don't.

We wish to make our publications widely accessible. Please contact us if you have any specific accessibility requirements.

First published by the Office of Qualifications and Examinations Regulation in 2013

This publication is also available on our website at www.ofqual.gov.uk

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

2nd Floor
Glendinning House
6 Murray Street
Belfast BT1 6DN

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346