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

Final Report

February 2014

Ofqual/14/5379

Contents

Foreword

This report presents the findings following our year-long study into the quality of marking of external exams in general qualifications in England. The study aimed to improve public understanding of how marking works, identify where current marking arrangements work well and recommend improvements in marking, where necessary.

This follows on from an interim report published in June 2013. This earlier report set out how marking works today and commented on significant developments in marking in the last decade.

Our findings show a complex and highly professional system, supported by expert examiners and improving technologies. The marking system in England is unusual in its scale, but it is well organised and tightly controlled, particularly at the most vulnerable points in the process. Fundamentally, we believe this is a system that people can have confidence in.

While marking is good, it can be better and we will be demanding more from exam boards in the future. Improvements include making better use of the potential offered by on-screen marking, as well as improving monitoring of traditional marking.

With 16 million scripts being marked every summer, mistakes do, and will, happen. Although there are few genuine mistakes, we cannot forget the impact they have on students and on confidence in marking. We will, therefore, require exam boards to offer a more transparent approach to marking, providing us with more data on the quality of their marking and sources of error in the system.

Of course, marking is not the end of the story for exam papers: schools and colleges can request an appeal through the enquiries about results and appeals processes.

These processes are coming under increasing pressure, and we want to change them so they are fit for the future. As a result, we will fundamentally re-design the entire appeals system in England. The new arrangements will be transparent, fair and robust enough to tell apart legitimate variations in marks from genuine marking errors. We aim to have the new system in place for 2015.

We also intend to make the marking system more professional. While, as a whole, examiners are experienced, it is vital that senior examiners in particular have the skills and assessment expertise to design the high-quality assessments necessary for reliable marking.

Professionalising marking also requires teachers to become more invested in the process, supported by their schools and colleges. Our research found some teachers

had a limited understanding of marking and held misconceptions about the system. We also found not all schools were supportive of examining. This is a public system for all, and everyone has a part to play.

1. Summary, key findings and next steps

Society trusts exam boards to make sure the results young people achieve are an accurate and fair reflection of their attainment. An underpinning requirement of this is that exams are marked accurately and reliably.

When stakes are high for students, and for schools and colleges, everyone needs to have confidence in the grades awarded to students and the marking that leads to those grades. An increasing minority of teachers tell us they do not have this confidence in marking.

In early 2013, we launched this review of the quality of marking of external exams in A levels, GCSEs and equivalent academic qualifications (collectively known as general qualifications[1]). The term 'quality of marking' is broad. For this review, quality of marking is defined as the accuracy and reliability of marking. This is to say students are given a mark that is as close to their correct, true score as possible, no matter who marked their work.

The three aims of our work were to:

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

2. identify where current arrangements work well (and where they do not);

3. recommend improvements, where they might be necessary.

We published an interim report called *Review of Quality of Marking in Exams in A levels, GCSEs and Other Academic Qualifications – Interim* in June 2013. This set out how marking works today and commented on significant developments in marking in the last decade.

In this report, we present the findings from our study and propose a set of improvements to the marking system. Due to the scale of the review, we accompany this report with eight supporting documents. These are:

- *Review of Quality of Marking in Exams in A Levels, GCSEs and Other Academic Qualifications - Findings from Survey of Examiners, May 2013*
- *Ofqual Quality of Marking - Qualitative Research Study*
- *Quality of Marking in General Qualifications - Survey of Teachers 2013*
- *Quality of Marking - Review of Literature on Item-level Marking Research*

[1] These include IGCSEs, International Baccalaureate (IB) Diploma, Pre-U Diploma and IGCE A levels.

- *Standardisation Methods, Mark Schemes, and Their Impact on Marking Reliability*
- *Review of Double Marking Research*
- *Review of Marking Internationally*
- *Quality of Marking: Description of the Marking Process Used in External Exams in General Qualifications.*

1.1 Key findings

Marking is a large-scale and complex exercise, but the system is well organised and tightly controlled, particularly at the most vulnerable points in the process. Exam boards' systems vary. We found no evidence that one system was better than another in principle, although isolated incidents of poor practice have been reported in some exam boards. Fundamentally, we believe people can have confidence in this system.

The marking system relies on 51,000 examiner roles.[2] Examiners are skilled and highly qualified – almost all have significant subject expertise and are experienced teachers, many with senior roles. They are confident, positive and are monitored increasingly responsively by exam boards.

The concept of valid assessment is central to the English exam system.[3] We cannot significantly improve the current reliability of exams without making them a less valid measure of skills and knowledge, or drastically increasing the amount of assessment students take. We reviewed available data from a small sample of GCSEs and A levels and found the sample to be at or above the minimum expected levels of reliability.

However, a significant weakness of the system is the lack of agreed metrics available to measure marking reliability (as opposed to qualification reliability) more specifically. We cannot currently compare the quality of marking of qualifications and subjects between exam boards. Until we can readily identify problem syllabuses using metrics, it will be difficult to improve wider confidence in marking.

Marking is not an exact science. While examiners can mark multiple-choice responses with precision, it is far harder to judge the quality of an essay response.

---

[2] Exam boards told us they use around 51,000 examiners, but we believe the total number of unique examiners working in the system to be around 34,000 (Ofqual, 2013a).

[3] We discuss the concepts of validity and reliability in section 2.

Extended response questions will always leave scope for legitimate differences of opinion between equally qualified and skilled examiners. Most exam boards recognise this legitimate variation by using marking tolerances during live marking.

Despite the strengths of the system, in a system of this scale mistakes do inevitably happen. There are only a small number of genuine marking errors. In summer 2013, just over 1 in 200 A level and GCSE certifications received a grade change as a result of the enquiries about results process. However, this is likely to over-estimate any rate of marking error, as four in five mark changes made after an enquiry about results were within the original marking tolerance and many were, therefore, likely to reflect legitimate variation in judgement by examiners.

Despite this, each year, several thousand mark changes are outside the original marking tolerance. Even though they affect a very small proportion of scripts, we cannot forget that any larger inconsistencies or marking errors can have significant consequences for the students involved. They can also have a real impact on school and teacher confidence.

We need a strong enquiries about results and appeals system to address errors and large inconsistences in marking. The current system does not enjoy the confidence of schools and colleges. It is both overly complex and conceptually at odds with how qualifications are marked initially. Moreover, we found schools and colleges were using it tactically due to pressures to deliver results, particularly at the top grades at A level and the C/D grade boundary at GCSE. The system was not designed to be used in this way and, while it is coping now, there is a risk it could bend under the pressure in time. We believe a full review is needed.

The marking system is in a steady period of transition, moving increasingly to online marking and training. These changes bring many positive developments, and more work can still be done to take further advantage of new technologies. Any period of change brings with it increased risks. There have been mistakes associated with the move online, but few considering the scale and significance of the transition.

Both teachers and examiners are positive about marking and generally have trust in its outcomes. However, an increasing minority of teachers and head teachers tell us they do not believe marking has been good enough in recent years, especially in GCSEs (Ipsos MORI, 2013). However, our review found marking arrangements for GCSEs and A levels were just as robust as those for equivalent academic qualifications; equivalent qualifications often use one and the same system.

There is a notable disconnect between how teachers perceive and understand the marking system and the reality of it in a number of areas. The better understanding teachers have of marking, the higher their confidence in it.

1.2 Next steps

Below, we set out a series of next steps we believe will further improve quality of marking and confidence in the marking system. We will lead this programme of improvements with exam boards and will do this in consultation with key stakeholder groups. Together, these changes will require us to review our regulatory tools. We are currently reviewing our *GCSE, GCE, Principal Learning and Project Code of Practice*[4] in light of these, and other, changes to the system.

1. Better monitoring and quantifying of the quality of marking of general qualifications

Previous research has found common indicators of marking quality are difficult to agree and set up (Tisi *et al.*, 2013).

We intend to develop a set of meaningful indicators that can be used, post-marking, to measure and monitor the quality of marking of general qualifications across the marking system, exam boards and qualification types. Schools and colleges will also be able to use these metrics to compare the quality of marking of different syllabuses. These measures will be published as part of a bigger suite of indicators, and they will be used to define acceptable levels of marking quality in different assessment types.

We will also require exam boards to publish information on significant marking errors in a transparent and timely manner.

2. Better data capture and feedback mechanisms to drive improvement

We will require exam boards to improve and formalise their data capture and feedback mechanisms to make sure the design of their assessments and marking processes are reviewed and refined on a regular basis. This should draw on item-level data gathered throughout live marking as well as evidence from the enquiries about results stage and following appeals. While some exam boards have working feedback loops in place, this is not universally a formalised process.

Exam boards should also be confident that their management information is robust enough to monitor effectively individual examiners and examining teams, as well as identifying wider marking issues. Broadly speaking, exam boards have commendable monitoring processes, but occasional large-scale marking mistakes show these are not effective at identifying marking problems at every exam board.

[4] www.ofqual.gov.uk/files/2011-05-27-code-of-practice-2011.pdf

3. Enquiries about results and appeals

We will fundamentally re-design the enquiries about results and appeals system in England. The new arrangements will be transparent, fair and robust enough to tell apart legitimate variations in marks and genuine marking errors.

This change will have a significant impact on exam boards' systems and processes. We aim to have a new system in place for the summer 2015 exam series.

4. Improving mark scheme design

We will lead on a formal programme of research to strengthen the evidence base on what makes a good mark scheme for students at all ability levels. This will focus particularly on those subjects and question types which can be most difficult to mark reliably, as well as how these mark schemes are applied to high performing students. We look forward to working with exam boards and other organisations on this.

Exam boards should improve mark scheme design using better data capture and feedback mechanisms, as discussed in point two, to find out how well mark schemes perform in practice.

The quality of mark schemes rests largely on the skills of senior examiners. Exam boards should continue to professionalise the roles of senior examiners, to make sure they all have the skills and expertise to design high-quality assessments.

5. Improving aspects of on-screen and traditional marking processes

Where exam boards move to on-screen marking systems, they must have the right infrastructure and processes in place. Any transition must take place at an appropriate pace for each exam board. Systems must be fully tested and all examiners trained appropriately to make the transition.

Exam boards should make better use of item-level marking through targeting questions at the examiners with the specialist knowledge to mark those items. This will let exam boards go further to eliminate sources of marking unreliability from the system.

Monitoring how examiners perform during traditional marking has limitations compared with on-screen examiner monitoring. Exam boards should look to transfer aspects of good practice from on-screen examiner monitoring to traditional marking, where this is feasible.

6. Better teacher understanding of and interaction with the exam system

The exam system is not just the responsibility of the exam boards and qualifications regulator. Schools, colleges and teachers should play their part in marking by actively improving their understanding of the marking system and supporting and participating in examining work.

Exam boards should circulate clear, up-to-date information on marking processes and systems to teachers, schools or colleges, and examiners.

During the course of this review, we identified some additional areas that might benefit from further research and evidence gathering. These include: the long-term impact of online standardisation on marking quality and examiner retention; the features of standardisation that improve marking reliability; the effectiveness of different models of marking quality assurance; the impact of formal training as part of making the senior examiner role more professional; and how double marking might be targeted effectively at certain assessment and item types.

2. The reliability of marking in England

The education system in England is unusual in having such a large number of general qualifications in different subjects, offered by a number of different qualification providers, and taken by so many students. It is also remarkable in the sense that almost all these exams need to be marked by expert examiners. This results in a marking system of considerable scale.

In summer 2012, 51,000 examiner roles[5] were needed to mark over 16 million exam papers taken by some 2 million students across seven exam boards.[6] In this context, the task of assuring quality of marking is a considerable one. But, where stakes are so high, exam boards must make sure marking is as accurate and reliable as it can ever be.

We ask a lot of final exams. A high-quality exam system must test students in both a valid and a reliable way. That is to say exams must measure what they set out to measure, and they must do so consistently.

Validity describes whether exam results are a good measure of what a student has learned. It makes sure we are testing the right knowledge and skills in the most fitting way. In this country, we value assessment validity over all else. However, validity is also underpinned by reliability. If reliability is not high enough, results are not a consistent measure of student performance and the assessment becomes meaningless. Put simply, reliability describes whether a student would have received the same result had he or she taken a different version of the exam, taken the exam on a different day, or had his or her work marked by a different examiner.

Validity and reliability are in careful balance, and it is important to be clear about what a marking system can ever reasonably deliver. Objective questions with unambiguous answers can be marked much more accurately and reliably than extended response questions. Examiners can mark multiple-choice responses with precision, but it is far harder to judge the quality of an essay response. When marking valid exams in English or history for example, examiners must make subjective judgements about student performance. Here, marking is not an exact science: a mark is a human judgement of a student's work and is only ever an approximation of his or her true score. Extended response questions will always

[5] Exam boards told us they use around 51,000 examiners, but we believe the total number of unique examiners working in the system to be around 34,000- see *Review of Quality of Marking in Exams in A levels, GCSEs and Other Academic Qualifications - Findings from Survey of Examiners, May 2013.*

[6] These are: AQA, Cambridge International Examinations (CIE), Council for the Curriculum, Examinations and Assessment (CCEA), the International Baccalaureate (IB), Pearson Edexcel, OCR and WJEC CBAC Limited.

leave scope for legitimate differences of opinion between equally qualified and skilled examiners (Tisi *et al.*, 2013).

Although multiple-choice and short-answer questions in exams can be marked more reliably, they are not always a valid means of assessing certain knowledge and skills in many subjects. We accept the lower levels of reliability associated with high-mark, stretching questions where we believe the question type is essential in assessing certain knowledge and skills at the appropriate level of demand. Introducing less valid means of assessment may narrow teaching and learning, and potentially reduce assessment demand.

It is also possible to improve reliability by increasing the amount of assessment each student must take, but the relationship is not a simple one. Significantly more assessment would be needed in each subject for only relatively small gains in reliability (Meadows and Billington, 2005).

For the reasons above, to have valid, demanding and manageable assessment in many subjects we must accept lower (but still satisfactory) levels of reliability. In this context, marking can never be totally consistent. We should not expect it to be. The qualifications system that best serves the education system in England would not result in near-perfect reliability. However, within these limitations, exam boards must make sure marking is as good as it can be and minimise genuine marking errors. Where mistakes do happen, exam boards must swiftly identify and remedy them.

Throughout this report, we use the term error to refer to genuine mistakes in marking.[7] Any variability in marks, large or small, which results from the inherent imprecision in marking is described as an inconsistency.

2.1 The reliability of A levels and GCSEs

Ultimately, marking quality is so essential because of its impact on qualification reliability and, consequently, the likelihood of students achieving the grade their performance merits. There are three major sources of unreliability in assessment, those that relate to: 1) the test – its structure, questions and mark scheme; 2) the students and their responses; and 3) the marking (Meadows and Billington, 2005).

The biggest source of unreliability is not marking, but the test itself (Wiliam, 2000; Bramley and Dhawan, 2012). As discussed previously, primarily this relates to the design of whole qualifications: the skills they test and the nature of the resulting tasks. It also relates to the quality of the design of individual questions and mark schemes. According to Meadows and Billington, the quality of a mark scheme is

[7] This is different to the concept of measurement error, which describes unavoidable imprecision in assessment.

central to an examiner's ability to mark well: "an unsatisfactory mark scheme can be the principal source of unreliable marking" (Meadows and Billington, 2005, p. 42).

We carried out extensive research into the reliability of general qualifications in our Reliability Programme (Ofqual, 2013b). As part of this, estimates of the reliability of a small sample of GCSEs and A levels were calculated. Appendix A presents the reliability coefficients[8] of AS chemistry, AS business studies and GCSE psychology at qualification level. AS chemistry and GCSE psychology had a composite qualification reliability of over 0.88 (above 0.92 in some cases). For AS business studies, the reliability was closer to 0.80.

Many experts have commented on the acceptable level of reliability in assessment. The general view is, when test scores are used to make important decisions about individuals in educational testing, a reliability coefficient lower than 0.80 would be considered as insufficient, between 0.80 and 0.90 sufficient, and above 0.90 good (Frisbie, 1988; Webb *et al.*, 2007; Evers *et al.*, 2010). The qualification-level reliability of the small sample of A levels and GCSEs, mentioned above, seems to be sufficient. However, as an assessment system, we have not formally agreed that this is indeed an acceptable level of reliability, or how such a level might vary depending on the nature of different assessment and question types. Agreeing these levels in a more informed manner would involve discussion with experts and collecting regular reliability data.

2.1.1 Marking reliability

Information on marking reliability of general qualifications is scarce. For one reason, studies of marking reliability need student responses to be marked at least twice. This can be difficult to organise, particularly during live exam marking, and often only relatively old studies exist (Murphy, 1978 and 1982, cited by Meadows and Billington, 2005). There is also considerable debate in the assessment community as to the most appropriate indicators of marking reliability (Tisi *et al.*, 2013). No common metrics of marking reliability have ever been collected and collated from exam boards.

The metrics currently commonly available in the system cannot readily be used for measuring marking quality across exam boards and qualification types. Enquiries about results are often used as a proxy indicator for a rate of marking error, but, as we outline in section 4, this is an imperfect metric. We also track teacher perceptions of marking. This is a helpful barometer of confidence, but perceptions do not always

[8] A reliability coefficient is a statistic used to quantify the consistency of test scores from repeated measurements. Reliability coefficients are used as estimates of the reliability of test scores and are usefully calculated as the correlations between two sets of scores.

reflect reality and can be affected by a range of factors besides marking. Live marking data is essential for monitoring examiner performance during live marking, though at present it does not readily give us any meaningful measures of overall marking quality at a system level.

This situation is not satisfactory. Exam boards should be able to demonstrate the quality of their outcomes, particularly where variations in practices both within and across boards exist. As a qualifications regulator, we must be able to monitor marking standards. The better the marking system and the more evidence available to show this, the more confidence schools, colleges and students will have in the results.

We now intend to develop a set of measures that can be used to monitor the reliability and accuracy of marking of general qualifications at a subject, exam board and qualification level. These measures must be robust, meaningful and should not, so far as is possible, introduce any perverse incentives into the exam system. Schools and colleges will be able to use these metrics to compare marking quality across subjects and exam boards.

In developing this suite of indicators, we will review good practice in other countries and industries as well as relevant assessment research. Measures might include rates of examiner agreement, examiner correlation and changes to marks and grades made as a result of a reformed enquiries about results system. These measures will be published as part of a bigger suite of indicators and we will use them to define acceptable levels of marking quality in different assessment types. For example, acceptable levels are likely to vary for extended response questions and low-tariff constrained questions.

2.1.2 What affects the reliability of marking?

Without meaningful metrics of marking quality (including marking reliability), we must instead identify which factors have the biggest impact on marking quality, and assess how these are currently performing in the English system. We know many factors influence the reliability and accuracy of exam marking. Studies frequently show that the single most crucial factor is the design and structure of the test and the items in it (Meadows and Billington, 2005; Tisi *et al.*, 2013). In England, we accept a compromise here. We could quickly improve marking reliability by only using constrained or multiple-choice questions. But we accept that wholly multiple-choice or constrained questions would not work for exams in many subjects and qualifications, and is not a viable option.

But we can influence other factors. High-quality assessment design is crucial in optimising marking reliability, no matter the style of assessment. Research has revealed an unsatisfactory mark scheme can be one of the principal sources of

unreliable marking (Meadows and Billington, 2005; Tisi *et al.*, 2013). Likewise, examiner expertise and experience are crucial to the marking of certain question types (Meadows and Billington, 2007; Suto *et al.,* 2011). The marking system in England is particularly strong here (see section 6). Examiner training, or standardisation, can also influence quality of marking, although the evidence on this is more mixed (Raikes *et al.*, 2004; Baird *et al.*, 2004).

Quality controls in marking have an impact on reliability through their efficiency in identifying and removing errant examiners. Finally, double marking[9] is also cited as a means to improve marking reliability. A feature of other marking systems around the world, double marking is not used in England in its truest sense. Here, some exam boards use double marking on a sample of scripts to monitor examiner performance, but the two marks are not combined to give an overall score. We recently completed a review of double marking, published in our supporting document *Review of Double Marking Research*.[10] This found a strong body of evidence from the 1940s to 1980s that double marking is more reliable than single marking. In contrast, studies carried out in the last 20 years were less compelling. These suggested that, while double marking does usually improve marking reliability, these gains are often smaller than expected, and would have a negligible impact on students achieving their true grade.

These gains must be weighed against the significant logistical and financial challenges of setting up double marking. If widely used in particular subjects, double marking would need twice the number of examiners. Even if recruiting these examiners were possible, introducing such a huge number of inexperienced examiners into the system at one time brings obvious risks to quality of marking. It would also have real implications for the cost of exams.

Given that double marking appears to yield only a small increase in marking reliability in today's qualifications, it may not justify the extra costs involved, particularly when other quality control methods may be more cost effective. In his paper on the marking reliability of GCSEs in maths and English, Newton (1996, p. 418) noted a "trade-off has to be made between reliability and cost-effectiveness: with the very large increase in examination costs that double marking would incur it would have to yield very much more reliable results than single marking to be considered appropriate".

It has been suggested that, given the additional cost of double marking, it should be targeted at exams "where genuine benefit can be demonstrated" (Brooks, 2004, p.

[9] In double marking, two examiners independently assess each candidate response. The final mark is the combination of the two examiners' separate marks. The combination of double marks to produce a final score acknowledges legitimate differences in opinion can exist between examiners.

[10] www.ofqual.gov.uk/files/2011-05-27-code-of-practice-2011.pdf

21). Further research could be done here to see if double marking can be more gainfully targeted at a small number of specific paper or item types.

Given the drivers of good quality of marking referred to in this section (including examiner experience, quality controls and standardisation), marking arrangements in England appear well designed and generally fit for purpose. We discuss specific aspects of the system, and these drivers of marking quality, in detail throughout this report.

3. The marking process – a system in transition

Between March and September 2013, we carried out detailed system mapping of the marking processes used by the seven exam boards providing general qualifications in England. This did not identify any significant issues in the marking system. Indeed, it indicated that the high-level marking and quality assurance processes in place across all general qualifications were sound. The systems are complex, multipart arrangements, but well organised and tightly controlled, particularly at the most vulnerable points in the marking process. They meet and often exceed regulatory requirements set out in the *GCSE, GCE, Principal Learning and Project Code of Practice*[11] and appear consistent with the practices of many of our international counterparts (Lamprianou, 2004), which we discuss in more detail in our supporting document *Review of Marking Internationally*.[12] While these high-level processes are robust, we are aware that there have been sporadic issues in implementing these processes in all of the major exam boards.

The overall marking process in place across the exam boards is broadly the same:

- The process begins with standardisation in which the principal examiner trains examiners and team leaders on how to apply the mark scheme.
- Examiners gain approval to begin live marking through marking a set of scripts or items to an acceptable standard.
- Once in live marking, examiners are monitored through taking samples of their work, either continuously (where they mark on-screen) or periodically (where they mark traditionally).
- After marking, any final checks are completed to confirm that marking is as free from error as it can be.

3.1 Variations in marking practice

The exact details of marking processes vary from exam board to exam board (and, therefore, qualification to qualification). It is difficult to isolate aspects of relatively good or poor practice in individual exam boards as individual quality controls are carefully balanced. What works well in one system might upset the equilibrium of another. In any case, in a qualifications market we are not concerned about

[11] www.ofqual.gov.uk/files/2011-05-27-code-of-practice-2011.pdf

[12] www.ofqual.gov.uk/documents/quality-of-marking-review-of-literature-on-item-level-marking-research

variations in marking practice, so long as this does not undermine quality. Standardising marking processes across exam boards is not desirable for its own sake, not least due to the stifling effect it has on innovation. Different practices can produce equally good outcomes, and we will focus on these outcomes through improved monitoring of metrics.

Below, we show just some of the variations in processes across the exam boards. We do this for the sake of transparency, and not as a judgement on which system may be preferable. We discuss the variations in marking practice in full in *Quality of Marking: Description of the Marking Process Used in External Exams in General Qualifications*.[13]

Generally, variations in marking practice are subtle. They exist at all stages of marking, from standardisation to post-marking data checks. Perhaps the biggest differences in practice can be seen in the monitoring of examiners during live marking, and we show some of these variations below.

In on-screen marking, exam boards predominantly use seed scripts or items[14] to monitor examiners. These seeds are essentially a set of test scripts or items. They have been pre-marked by a senior examiner and appear at random in an examiner's marking allocation. At least 1 in 20 scripts or items marked by an examiner is likely to be a seed, (1 in 10 in the IB). AQA and WJEC also require examiners to pass a number of seeds when they log into the online system each day.

Examiners must mark seeds to a given standard. If their marking falls below this standard, they are monitored more closely or temporarily stopped from marking and given feedback. If marking doesn't improve, examiners are stopped permanently from marking. Their scripts are re-marked by another examiner, either in full or from the last point at which they were known to be marking accurately.

Exam boards vary in their seeding strategy: OCR and WJEC may vary the seed rate depending on the nature of the paper being marked; CIE and the IB vary the seed rate by individual examiner if necessary; and Pearson Edexcel varies seed rates by examiner type (with different rates for general markers[15] and examiners). Different exam boards supplement seeding with other sampling techniques. For example,

---

[13] www.ofqual.gov.uk/documents/quality-of-marking-description-of-the-marking-process-used-in-external-exams-in-general-qualifications

[14] A item is an individual question or groups of related questions.

[15] General markers mark simple, highly constrained questions with clearly defined answers. They are 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.

either in addition to or instead of seeding, AQA and WJEC double mark a sample of 5 to 10 per cent of items in subjective papers to make sure examiners are marking within an acceptable tolerance of each other. All exam boards expect principal examiners and team leaders to carry out extra spot checking, or back reading, of marking alongside these checks. Generally, this is left to the discretion of principal examiners and team leaders. However, in Pearson Edexcel there are guidelines around the amount of back reading that must be completed.

Sampling to check marking accuracy also takes place in traditional marking. Unlike sampling for on-screen marking, it cannot be continuous for logistical reasons, and the frequency, timing and size of exam boards' samples vary. For example, three exam boards (CIE, OCR and Pearson Edexcel) take two samples of examiners' marking during live marking and three exam boards (AQA, the IB and WJEC) take one sample. In traditional marking, common pre-marked scripts are not routinely used. Instead, examiners select scripts from their allocation and send a batch to their supervising examiner for review. Exam boards vary in how much discretion examiners have to select their samples: some exam boards (such as the IB) request specific scripts. Others let the examiner choose to submit any scripts they wish.

Whether marking on-screen or traditionally, examiners must pass sampling checks by marking a proportion of their scripts within either a set marking tolerance or an adjacency value.[16] In most cases, this tolerance recognises there can be legitimate differences in professional judgement between experienced examiners. The size and nature of marking tolerances vary across exam boards. The IB Diploma has the widest script-level marking tolerances. In English literature IB Diploma exams, marking tolerances are between 12 and 15 per cent of the total raw mark available for the unit. For biology, they are between 10 and 11 per cent. In contrast, tolerances at CIE for the Pre-U Diploma and International A level are far narrower: 4 per cent for English literature and under 2 per cent for biology.

Exam boards use these marking tolerances differently, and they trigger different interventions. For example although the IB's tolerances are widest, marking seeds outside tolerance triggers automatic suspension from marking. Where tolerances are much narrower (CIE and OCR), marking outside tolerance initially triggers the closer monitoring of an examiner – it does not necessarily indicate any sense of error (although this depends on the type of unit in question). These are clearly rather different interpretations of the word tolerance.

[16] For both traditional and on-screen marking, Pearson Edexcel uses adjacency values, which are not recognition of any acceptable level of variation. Instead, they are used as a flag to identify examiners who may need closer scrutiny and/or corrective action.

The examples above illustrate there are variations between the levels and types of sampling, the criteria for failing sampling checks and the action a failure triggers. These three variables are interdependent and, together with the size of the marking tolerance, they affect the likelihood of identifying an aberrant examiner. Changing any one of these factors affects the effectiveness of sampling in identifying underperforming examiners. Evidence in this area is limited and could benefit from more research to evaluate which controls should be set, at what level and in what combination to identify underperforming examiners most effectively.

3.2 Technological developments in marking

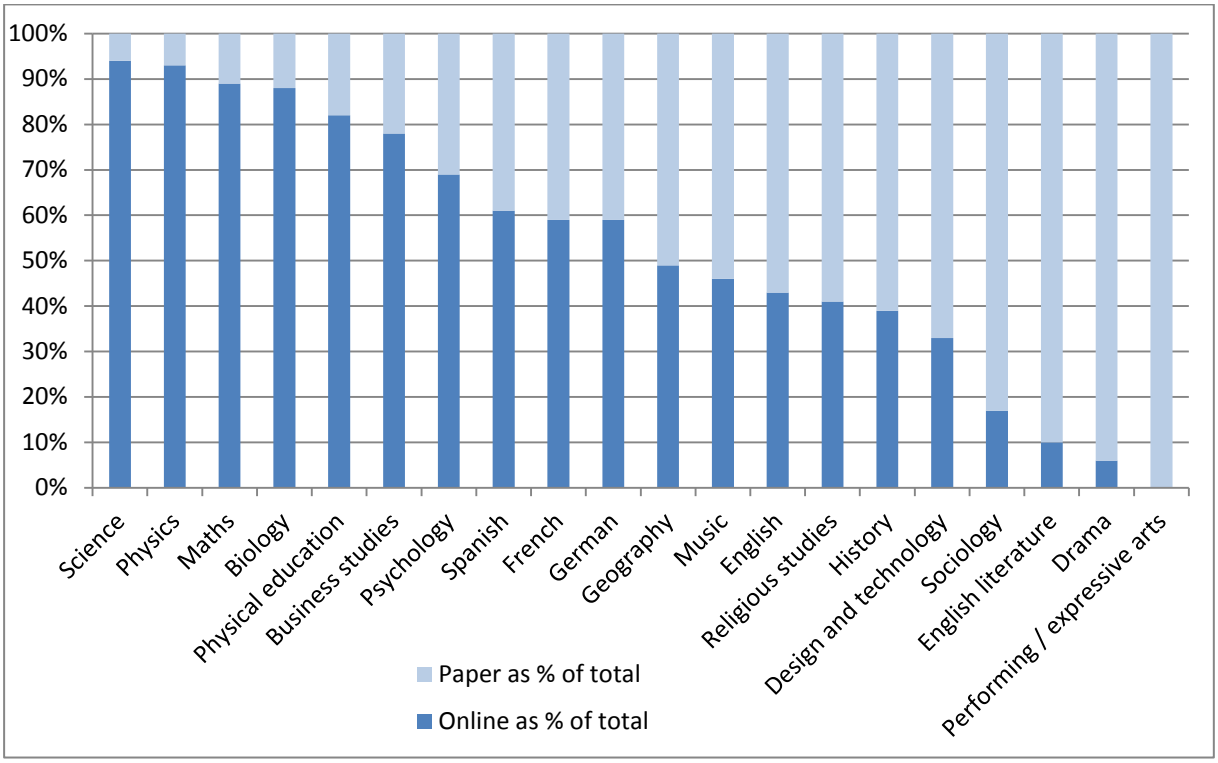
3.2.1 Prevalence of on-screen marking

The marking system is currently in a state of profound transition from traditional pen-and-paper marking and face-to-face training to on-screen marking and training. This has shaped all aspects of the marking process, from standardisation through to post-marking checks. As discussed in our supporting document *Review of Marking Internationally*,[17] England is by no means alone in adopting such new technologies. Many jurisdictions – including Hong Kong, the Republic of Korea, New South Wales (Australia) and Massachusetts (USA) – also use on-screen marking in their high-stakes assessments. Where this is the case, many also use item-level marking.

Overall, two thirds of marking is now carried out on screen. In summer 2012, this ranged from 88 per cent at Pearson Edexcel to 13 per cent at WJEC (Ofqual, 2013c). On-screen marking is most widely used for exams with shorter, more constrained questions. Essay questions are less likely to be marked on screen. Figure 1 shows that more objective subjects such as science and maths are far more likely to be marked on-screen than subjective disciplines such as English, drama and history.

[17] www.ofqual.gov.uk/documents/review-of-marking-internationally

Figure 1: Scripts marked traditionally and on-screen by subject, summer 2012



We believe these developments are positive. As well as improving the security and speed of marking, on-screen marking allows more sophisticated examiner monitoring. This is likely to improve marking quality through better identification and removal of aberrant examiners. However, any process of transition is not without its risks.

Over the last three years we have been notified by exam boards of three clear instances of marking mistakes affecting a group of candidates that were directly associated with exam boards' transition from traditional to on-screen marking.

When mistakes like this occur, we agree with exam boards their approach to remedy for students affected; recognising that when students have already acted on their results it could be inequitable to downgrade them.

In summer 2010, AQA experienced problems with the transition to on-screen marking for unconstrained, essay questions[18] in a range of GCSE and A level subjects. This resulted in some 3,340 students receiving the wrong marks and 622 incorrect grades being issued. AQA acted promptly and set in place an action plan to improve the robustness of its marking process. There have been no further transition incidents reported by AQA.

[18] In these cases, candidates normally write their response in a separate generic answer booklet, where the response area for each question is not pre-defined.

In September 2012, WJEC reported errors in the marking of two computer-marked items for GCSE Spanish Listening. As a result of human error in the process for deciding what each answer was worth, a number of answers were credited with fewer marks than they deserved.

The errors affected approximately 3,600 candidates and resulted in 193 candidates gaining lower qualification grades than they should have. WJEC acted promptly to correct the errors and re-issue students' results. We required WJEC to carry out a full and thorough review of its procedures for on-screen marking and make improvements to the training for staff.

Mistakes may not be immediately apparent. In late 2013 OCR notified us of mistakes associated with the introduction of on-screen marking of four A level units in English and history in the summer 2013 series, discovered well after the event. Investigations are ongoing but the total number of affected students is likely to be low.

Given the potential for the types of issues above, all exam boards must have the right infrastructure and processes in place to support transitions to online systems. Transitions must take place at an appropriate pace. All systems must be fully tested and all examiners appropriately trained to make a transition. Where on-screen marking is set up properly, the benefits are considerable, but any transition process must be managed and monitored closely.

3.2.2 On-screen and traditional examiner monitoring

For all the benefits of on-screen marking, some exam boards still favour traditional marking. This is perfectly acceptable. Paper-based marking has been used in exams for decades and has many advantages, including high examiner engagement. Some exam boards believe marking essay questions on-screen can be cognitively more challenging for examiners. While research confirms this may be true, this does not affect the accuracy or reliability of marking, either for short-answer or essay questions (Johnson *et al.*, 2010; Johnson *et al.*, 2012). On-screen and traditional marking are two equally valid methods of marking. However, it is apparent they do not currently deliver the same level of examiner monitoring.

In live on-screen marking, sampling is both continuous and blind. Research has shown marking consistency can be over-estimated when the second examiner can see the first examiner's marks (Tisi *et al.*, 2013). Blind re-marking is, therefore, accepted to be the most effective way of detecting examiner inconsistency through a sampling approach (Billington, 2012). Seeding is a form of blind re-marking. As the definitive mark for each seed is determined before live marking, a senior examiner's judgement of examiner work is not technically a re-mark, although the principle is the same. Both examiners and senior examiners give marks to a clean script or item

entirely independently, before a comparison of their marking is made. Therefore, a form of continuous, blind re-marking of a sample of examiner work is now the norm during live marking for most exams in England (although not at later enquiries about results and appeals stages).

The monitoring of traditional marking has remained broadly the same for decades. It is more limited than on-screen monitoring, mainly due to the logistical difficulties in moving physical scripts between examiners in a tight marking window. Exam boards generally take one or two samples of examiner work during live marking. Studies show there can be changes in how severely individual examiners mark over time (Myford and Wolfe, 2009; Baird *et al.*, 2012). Therefore, such an isolated sampling process might fail to spot changes in marking behaviour over time (Tisi *et al.*, 2013). Another limitation of traditional monitoring is that the senior examiner can see the marks and annotations made by the first examiner. Re-marking is not blind and that means it may under-estimate examiner inconsistency (Billington, 2012).

Given the research above, **exam boards should consider how they can apply some of the techniques for monitoring examiners on-screen to traditional marking**. It could be possible to incorporate some blind re-marking into traditional sampling methods by giving examiners additional common pre-marked scripts. It may also be feasible to increase the number of live marking samples, and tighten the parameters for selecting scripts to include in samples. Exam boards should investigate the impact of such changes on quality of marking, alongside any other potential strategies for improving the monitoring of traditional marking.

3.3 Item-level marking

On-screen marking opens up possibilities for other changes to marking processes. One of the more significant of these is the move to item-level marking. In summer 2012, just under half of scripts were marked at item-level. In item-level marking, a scanned script is split up into individual questions (or groups of related questions), which are marked by different examiners. AQA, Pearson Edexcel and WJEC all use item-level marking for their on-screen marking. The IB is trialling item-level marking across six subjects in 2014, and CIE is also considering introducing it for science subjects. OCR does not use item-level marking in its truest sense but allows examiners to mark their allocation of whole scripts item by item.

Item-level marking has many potential benefits that, in theory at least, could improve marking reliability and accuracy. These include:

- Reducing the effect of biases caused by student responses to the rest of the exam paper. This is known as the halo effect and is eliminated in item-level marking (Spear, 1996).

- Reducing the influence of a single examiner on an exam script. Variations in marking often cancel each other out. That is, for each question that is over-marked there is likely to be one that is under-marked (Pinot de Moira, 2011).
- Enabling questions to go to examiners with the appropriate level of expertise.

These theoretical benefits are supported by a small body of research, which we discuss in our supporting document *Review of Literature on Item-level Marking Research*.[19] The research shows item-level marking is at least as reliable as whole-script marking, and under some conditions is likely to be more reliable (Wheadon and Pinot de Moira, 2013; Pearson Edexcel, 2003 and 2004; Black and Curcin, in preparation). For example, Wheadon and Pinot de Moira (2013) analysed marking data from two AQA A level geography units that switched between whole-script marking and item-level marking over the course of three years. The study found item-level marking appeared to improve the reliability of marking, in particular for the highest performing students.

Black and Curcin (in preparation) found evidence of the halo effect at work in whole-script marking, which was not there in item-level marking. Item-level marking also removed the most extreme differences between a student's true grade and the grade awarded through whole-script marking, although it did not yield substantial advantages in terms of students achieving the correct grade.

For all its potential benefits, some teachers and stakeholder groups have concerns about item-level marking. The Royal Historical Society tells us teachers worry that item-level marking makes it more difficult for examiners to "take an effective overall view of a script" (Dodd, 2014, p. 31). We believe these fears are misplaced. When marking a paper, each examiner is trained to apply a mark scheme consistently to every question across a script. To ensure fair and reliable exams, examiners must apply these mark schemes consistently. To introduce an undefined sense of holistic judgement to the marking process is to introduce a source of unreliability. Item-level marking removes this risk just as it removes the halo effect. We believe this is one of its benefits.

The Royal Historical Society, Association of School and College Leaders) and English Association all recognise item-level marking's potential in targeting examiner expertise more effectively. In history or English subjects, a single paper can cover many different historical periods or set texts. We know that topic or text unfamiliarity can be a source of error in certain subjects where a wide breadth of content is assessed (see section 4). Allocating specific questions "according to marker

[19] www.ofqual.gov.uk/documents/quality-of-marking-review-of-literature-on-item-level-marking-research

preference/expertise could result in more accurate assessment" (Dodd, 2014, p. 31). Exam boards in England have yet to exploit this intelligent targeting of items to examiners. **We believe exam boards should make better use of item-level marking through targeting questions at examiners with the specialist knowledge to best mark those items. This will allow exam boards to go further to eliminate sources of marking unreliability from the system.**

3.4 Online standardisation

Standardisation makes sure all examiners are fully competent in applying a mark scheme before they begin marking. Traditionally, standardisation was carried out in face-to-face meetings chaired by principal examiners. More and more often, however, standardisation is delivered online, with examiners working through examples of student scripts on-screen. In summer 2013, online standardisation overtook face-to-face meetings as the main form of standardisation in general qualifications, accounting for 39 per cent of all units. As shown in figure 2, the amount of online standardisation varies significantly by exam board.

As with on-screen marking, online standardisation is more common for papers with lower tariff, constrained items. In summer 2013, subjects such as maths (54 per cent) and biology (52 per cent) were most likely to be standardised online. In contrast, more subjective disciplines such as drama (75 per cent) and English language (67 per cent) were likely to be standardised face-to-face. Face-to-face standardisation was more common in A levels than GCSEs. The IB Diploma was the only qualification in which face-to-face standardisation was limited to some senior examiners.

Figure 2: Methods of standardisation used by exam boards in summer 2013

Exam board	% online standardisation	% traditional standardisation	% webinar standardisation[20]	% no standardisation required or other techniques used[21]
OCR	83%	17%	0%	0%
AQA	55%	34%	0%	11%
CIE	35%	44%	0%	21%
Pearson Edexcel	27.5%	25%	35%	12.5%
The IB	21%	0%	12%	67%
WJEC	3%	97%	0%	0%
CCEA	0%	100%	0%	0%
Average	39%	38%	7%	16%

As in England, standardisation practice varies worldwide. Some jurisdictions use face-to-face standardisation, whereas others prefer online approaches (Boyle *et al.*, 2014). International research into examiner standardisation training found examiner training was "one of the most important tools… to improve agreement" (Center for Educator Compensation Reform, 2012, p. 15). Exactly which aspects of this training have the most impact on marking quality is unclear. However, studies agree that no matter how good the training, it cannot remove marking inconsistencies completely (Haladyna *et al*., 2013).

A literature review on the effectiveness of online standardisation found it could provide marking at least as high quality as traditional methods, although, admittedly, research is limited (Boyle *et al.*, 2014). Wolfe, Matthews and Vickers (2010) studied examiner performance on a state-wide writing assessment in the USA, using three types of standardisation training, including face-to-face and online. They found quality was highest in an online group. In England, Chamberlain and Taylor (2010) found online standardisation was as effective as a face-to-face meeting, even in a subjective paper such as A level history. Knoch, Read and von Randow (2007, cited by Boyle *et al.*, 2014) discovered that online training was slightly more successful at

[20] Webinars use web platforms to host virtual standardisation meetings. They attempt to simulate some of the discussion between examiners fostered in traditional standardisation.
[21] The remainder of units either needed no standardisation (the principal examiner completed all the marking) or used other techniques. Other techniques might include moderation, where an examiner's marking is brought in line with the standard using a statistical approach.

achieving marking consistency when training examiners of writing assessments at a New Zealand University.

One shortcoming of the studies to date is they do not take into account any residual effect of traditional standardisation on marking quality. We know the examining workforce in England is highly experienced. It is possible that previous good practice has prepared examiners well enough to mitigate any weaknesses of online standardisation. Over time, this embedded understanding could be eroded, to the detriment of marking quality. We expect exam boards to monitor this closely.

3.4.1 Perceptions of online standardisation

In spite of the encouraging findings above, online standardisation seems to be highly unpopular with examiners. This dislike of online standardisation is not down to the technology more widely. In a series of in-depth interviews with some 50 examiners, many were positive about on-screen marking, but they also voiced an overwhelming opposition to online standardisation. They claimed online standardisation undermined their ability to mark accurately and consistently, notably in subjective paper types. They did not believe online standardisation was as effective at helping examiners to understand the mark scheme at the profound level they thought necessary. Examiners can also find it difficult to absorb every piece of written guidance that comes with online standardisation (Oxygen, 2014).

In contrast to the criticisms above, other evidence paints a more balanced picture. Eighty-five per cent of the some 10,000 examiners who responded to our 2013 examiner survey agreed with “I receive sufficient briefing about a paper and mark scheme before I begin my marking for each exam”. While the move to online standardisation is unpopular, standardisation is clearly perceived to work effectively for most examiners. Nonetheless, when survey respondents were asked how the marking system might be improved, they most frequently suggested a return to face-to-face standardisation (Ofqual, 2013c).

The combined evidence paints a mixed picture of the impact of online standardisation on marking quality. Examiners’ strength of feeling towards online standardisation belied the fact they generally felt well briefed for marking under any system. What is evident is they felt better prepared and more confident when standardised face-to-face. However, there is no established link between higher confidence and better marking. Some studies found perceptions of training were actually a poor indication of its effectiveness (Boyle *et al.*, 2014). This is illustrated starkly in a study by Knoch (2011, cited by Boyle *et al.*, 2014), who found a disassociation between perceptions of the training, and its impact. In this study, examiners who benefited from training didn’t like it, whereas those whose marking was not improved by the training, did.

3.4.2 Online standardisation and examiner engagement

Examiners' second major criticism of online standardisation was it made marking less rewarding. They commented the new system felt isolating and took away from the enjoyment of the examiner role. Senior examiners raised concerns about the effect of this on retaining and recruiting the best examiners, and the impact this could have on future marking quality. They believed many examiners put great value on the social and professional networking aspects of traditional standardisation. Online standardisation delivers fewer rewards from a professional development standpoint and reduces this incentive to examine (Oxygen, 2014).

This concept of examiner engagement is not a new one. Some of the earliest studies of online standardisation discussed the notion of communities of practice in examining (or shared understanding between professionals), and warned these communities might be eroded by online standardisation (Meadows and Billington, 2005). However, more recent studies called into question the importance of communities of practice in ensuring marking reliability (Baird *et al.*, 2004).

That is not to say communities of practice do not have an important place in marking. This social aspect of standardisation could be important in maintaining teachers' engagement with the exam system (Boyle *et al.*, 2014). It could also help to retain experienced examiners. There has been little research on how online standardisation impacts examiner retention, and reasons for examiner drop-out are rarely recorded.

At present, the evidence available does not suggest that online standardisation is a real threat to quality of marking. Exam boards should continue to monitor this. But it does appear to lessen examiner confidence and their enjoyment of the role. All exam boards must make sure the increasing use of technology does not lead to a shortage of examiners, more so than ever during this period of A level and GCSE reform.

4. Marking errors and inconsistencies

Sometimes, marking does go wrong. It is inevitable in any complex system of this scale that mistakes will happen, no matter how extensive the systems and controls that aim to prevent this. It is certainly not unique to the marking system in England (Lamprianou, 2004). But we must not be complacent about this. It is critical that genuine errors are kept to a minimum, and there is a robust system in place for challenging marking when errors have happened. This is the enquiries about results and appeals system. We discuss this system in detail in section 8.

While errors clearly do happen in marking, we must tell apart genuine mistakes and inconsistencies due to justifiable differences of opinion between equally skilled examiners. The former we can, and must, address. The latter we need to accept as an inevitable and quite legitimate part of a valid assessment system, although efforts should always be made to reduce its magnitude. Marking errors and inconsistences between examiners affect public confidence in marking. Experience of a larger error can be particularly damaging, and have a long-term impact on how a teacher views marking (Oxygen, 2014; Dodd, 2014).

4.1 The prevalence of marking errors

At present, there is no metric that gives an error rate for the marking of general qualifications in England. As discussed in section 2, this is not satisfactory and we will remedy this. In the meantime, we rely on proxy indicators and qualitative data to identify incidences and patterns of marking error. At present, the closest proxy of error we have is the rate of enquiries about results submissions and grade or mark changes. The enquiries about results system allows schools or colleges to request a review of the marking of a student's script where they believe an error has been made.[22] The information below draws on the most recent published data for A levels and GCSEs from summer 2013 (Ofqual, 2013d), as well as data for all general qualifications from summer 2012 specifically collected for this review.

After the summer 2013 exam series, exam boards received 301,250 enquiries about results for external assessments in GCSEs and A levels. This represents 2.3 per cent of all scripts, an increase from 1.9 per cent in summer 2012. While the number of enquiries has increased, the number of enquiries resulting in a qualification grade change has remained stable over the last five years. In 2013, 16.5 per cent of enquiries led to a qualification grade change. This represented 0.6 per cent of all certifications in the UK. On this evidence, we might conclude that a little over 1 in 200 exam papers contained a marking error or inconsistency.

[22] If the investigation shows marking or processing errors have been made and the candidate's result is incorrect, the awarding organisation will adjust the mark. In some cases, this may affect the overall qualification grade, which will then also be adjusted.

However, research shows enquiries about results data is not always a good indicator of error for various reasons. Firstly, schools will not submit an enquiry for every marking error (particularly when the mark is higher than expected). There are limitations caused by the enquiries about results process itself (see section 8), as well as the behaviour of some schools or colleges in speculatively submitting enquiries for students whose marks are just below key grade boundaries. Moreover, enquiries about results do not distinguish between genuine errors and acceptable subjectivity in marking.

Mark changes made as a result of an enquiry are usually small. In summer 2012, the average mark change in external exams was between 1 and 2 marks. Given that schools and colleges overwhelmingly submit enquiries just below grade boundaries, many of these small changes result in a unit or qualification grade change. Mark changes to external exams led to 40,400 qualification grade changes across all general qualifications in summer 2012. Ninety-nine per cent of these were changes of one grade. Just 423 students received a grade change of two or more grades.[23]

When exam boards review the marking of external scripts during the enquiries about results process, they do not apply a formal numerical marking tolerance (with the exception of CIE).[24] No matter how small the mark change identified, a student's mark will be revised. Data from summer 2012 found 79 per cent of the mark changes resulting from enquiries about results in A levels, GCSEs and IGCSEs were within the marking tolerance used in live marking.[25] While some in-tolerance mark changes may still involve a marking mistake, it is feasible that the majority of mark changes are due to legitimate differences in opinion between examiners.

This is significant. It suggests that marking errors (and large inconsistences) could be up to five times less prevalent than reported. Public perceptions are of rising numbers of errors, when in actual fact, many mark or grade changes made as a result of an enquiry only reflect the inevitable variation in marking between examiners, rather than marking mistakes. Nonetheless, we should still acknowledge that in summer 2012, of the 124,800 mark changes made as a result of enquiries about results in A levels, GCSEs and IGCSEs, some 26,000 were outside the original marking tolerance.[26] This is a small number in a system of 16 million exam scripts,

---

[23] This data refers to A levels, GCSEs, the IB Diploma and IGCSEs (with the exception of those provided by CIE).

[24] However, a tolerance is applied in the re-moderation of enquiries about results involving internal assessment.

[25] Data provided by Pearson Edexcel, AQA and OCR. For Pearson Edexcel, adjacency values are used in place of tolerances.

[26] Data excludes IGCSEs provided by CIE.

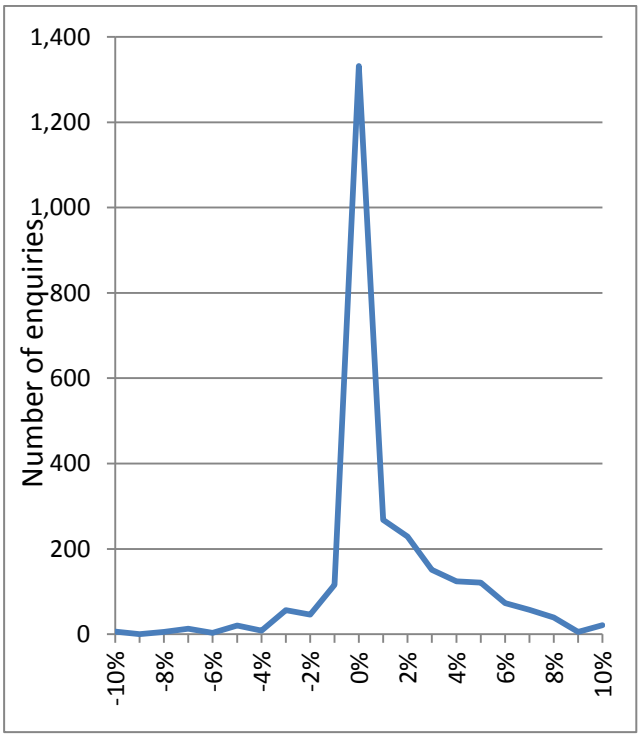
but it is significant enough that every school or college may experience one or more of these mark changes each year.

To illustrate the size and distribution of mark changes, we give two case studies below. The subjects were selected either because they had above-average rates of enquiries about results unit grade changes (geography) or they were subject to strong criticisms from teachers over marking (French).

4.1.1 Mark changes in French and geography – summer 2012

The data below shows enquiries about results mark changes in A levels and equivalent qualifications (the IB Diploma, Cambridge Pre-U and Pre-U Diploma, and International A levels). For both geography and French, the average mark change was less than one mark. Figures 3 and 4 present the mark changes for both subjects across all units as a percentage of the total marks available for the unit. The percentage mark changes for French and geography are low, with no overall changes to almost half the scripts. In both cases, over two thirds of enquiries about results led to a mark change of 2 per cent or less of the raw marks available for the unit (65 per cent in geography and 73 per cent in French).

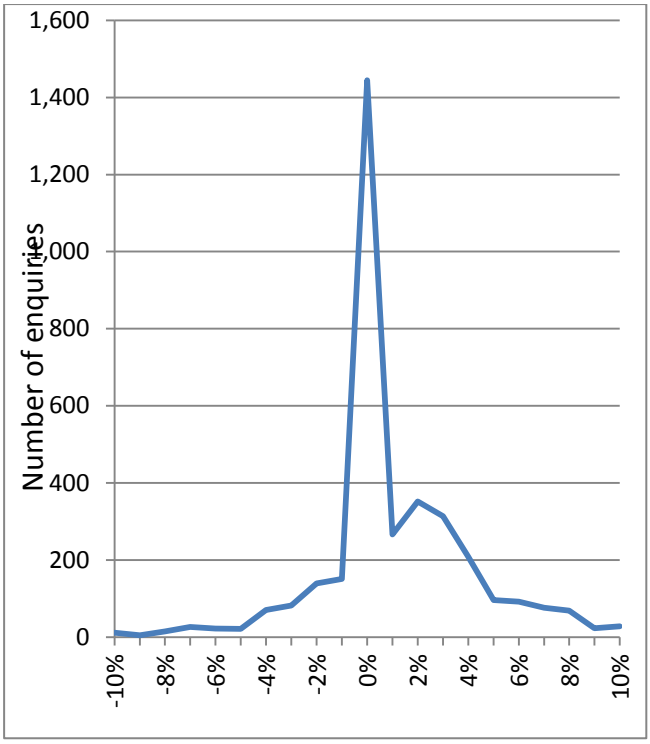
Figure 3: Geography mark changes as a proportion of the total raw marks available per unit



Geography A level and equivalent

- 4,307 enquiries about results in summer 2012 (3 per cent of all scripts);
- forty per cent of enquiries about results led to no mark change;
- eighty-seven per cent of mark changes were within 5 per cent of the total raw marks available for the unit;
- one per cent of mark changes were 10 per cent or more of the total raw marks available for the unit;
- 769 qualification grade changes were made;
- sixty-six per cent of these were the result of mark changes within the original marking tolerance.

Figure 4: French mark changes as a proportion of the total raw marks available per unit



French A level and equivalent

- 3,270 enquiries about results in summer 2012 (5 per cent of all scripts);
- forty-nine per cent of enquiries about results led to no mark change;
- ninety-one per cent of mark changes were within 5 per cent of the total raw marks available for the unit;
- one per cent of mark changes were 10 per cent or more of the total raw marks available for the unit;
- 428 qualification grade changes were made;
- fifty-nine per cent of these were the result of mark changes within the original marking tolerance.

In 2012, mark changes in units for geography were often greater than for French, although these mark changes were still within a small range. Both subjects also showed low rates of large mark changes – less than 1 per cent of mark changes were 10 per cent or more of the raw marks available for the unit. Perhaps most significantly, around 60 per cent of qualification grade changes for both subjects were within the original marking tolerance. In geography units, just a third of qualification grade changes were the result of a mark change outside this tolerance.

4.2 Reasons for mark changes

Some exam boards do not routinely record reasons for mark or grade changes as a result of an enquiry about results. We believe better recording of this is important for improving the feedback loop within exam boards and making it more likely to prevent errors in future, and for understanding the causes of marking error at a system level. Greater transparency about larger scale errors also lets other exam boards learn lessons from such incidents. **We will require exam boards to publish information on significant marking errors in a transparent and timely manner.**

Exam boards have to notify us of events that could have an adverse effect on students. These can cover a wide range of incidents, but from a marking perspective, they include cases where they have issued incorrect results or certificates to

students. Between March 2011 and May 2013, we received 33 notifications of marking errors. We take these reports, and any other instances of marking error, extremely seriously. Over the past five years, we have investigated sporadic but significant marking errors at a number of the exam boards providing general qualifications in England. While human error is inevitable in a marking system of this size, more widespread, systemic errors are not acceptable. Where we receive reports of such errors, we will consider taking regulatory action against the exam boards involved.

Based on our event notifications, the most frequently reported type of marking error was mark scheme error, accounting for around one in three event notifications. These included: late changes to papers not being carried through to mark schemes; correct answers not being included where multiple answers are allowed; and multiple-choice mark schemes having an incorrect answer key. The second most common error was examiner error. This tended to happen where an examiner had not applied a mark scheme correctly and it had not been picked up until after results were issued.

IT and system errors have also been responsible for marking errors. Usually, the marking has been completed correctly, but a system issue has caused an error in results. Linked to this are cases involving on-screen marking where some parts of a student response have been left unmarked. Finally, there are clerical errors resulting from mistakes in the manual totalling or transcription of marks. This pattern of marking errors is broadly similar to the types of errors experienced internationally (Lamprianou, 2004).

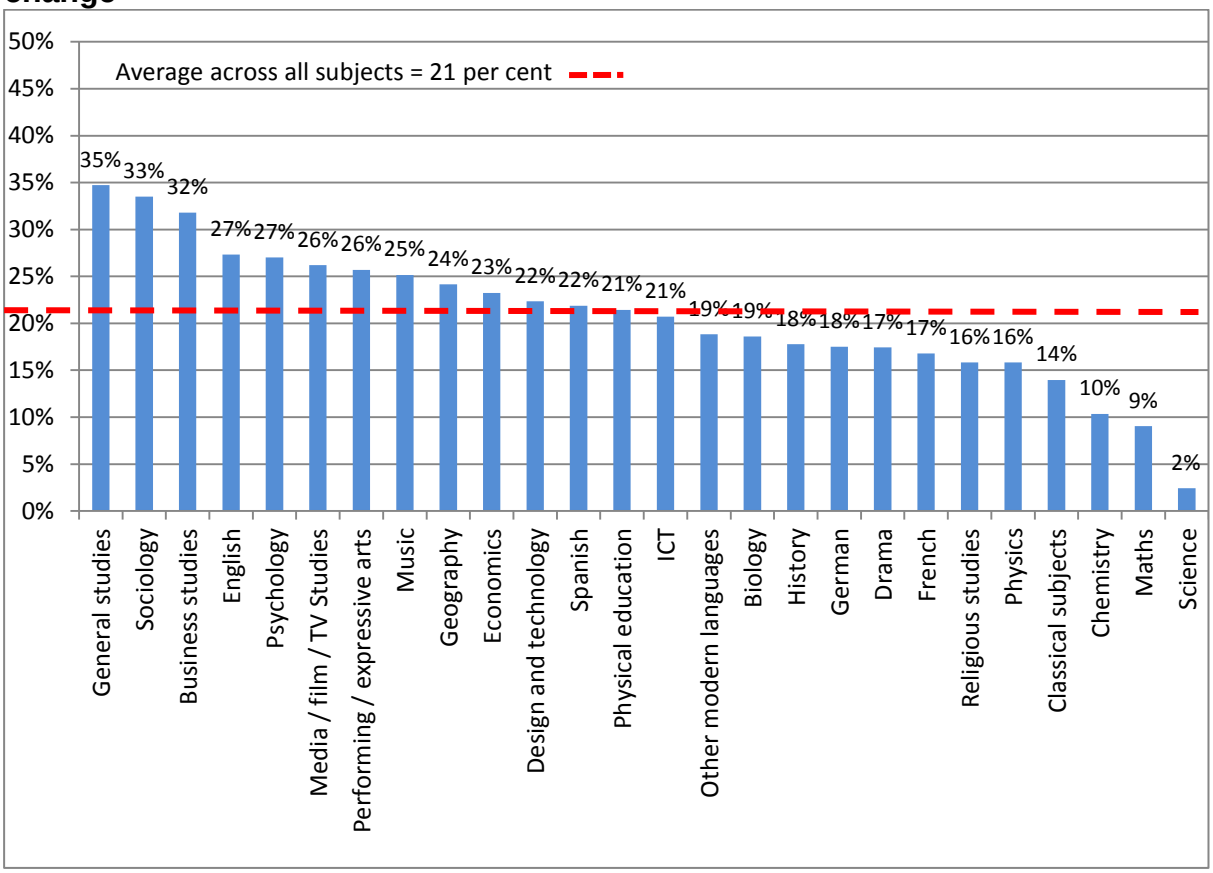
4.3 Where are marking errors most likely to happen?

Below, we have analysed patterns of enquiries about results by subject for A levels and GCSEs in summer 2012. Given the high percentage of mark changes that were within the original marking tolerance, this is possibly more likely to tell us about inherent subjectivity in subjects rather than marking mistakes.

At A level, enquiries about results were most likely in music, classical subjects, modern foreign languages (French and Spanish), economics and history. There was a general trend whereby subjective disciplines were more likely to receive an enquiry about results than subjects such as science and maths. There were some exceptions to this – art and design, critical thinking, and sociology all had low rates of enquiries about results per entry. At GCSE, there was a similar pattern, with more subjective disciplines more likely to receive an enquiry about results. However, at GCSE, the core subjects of maths and English were more likely to be subject to an enquiry about results than at A level.

The rates of unit grade changes resulting from enquiries about results are shown in figures 5 and 6.

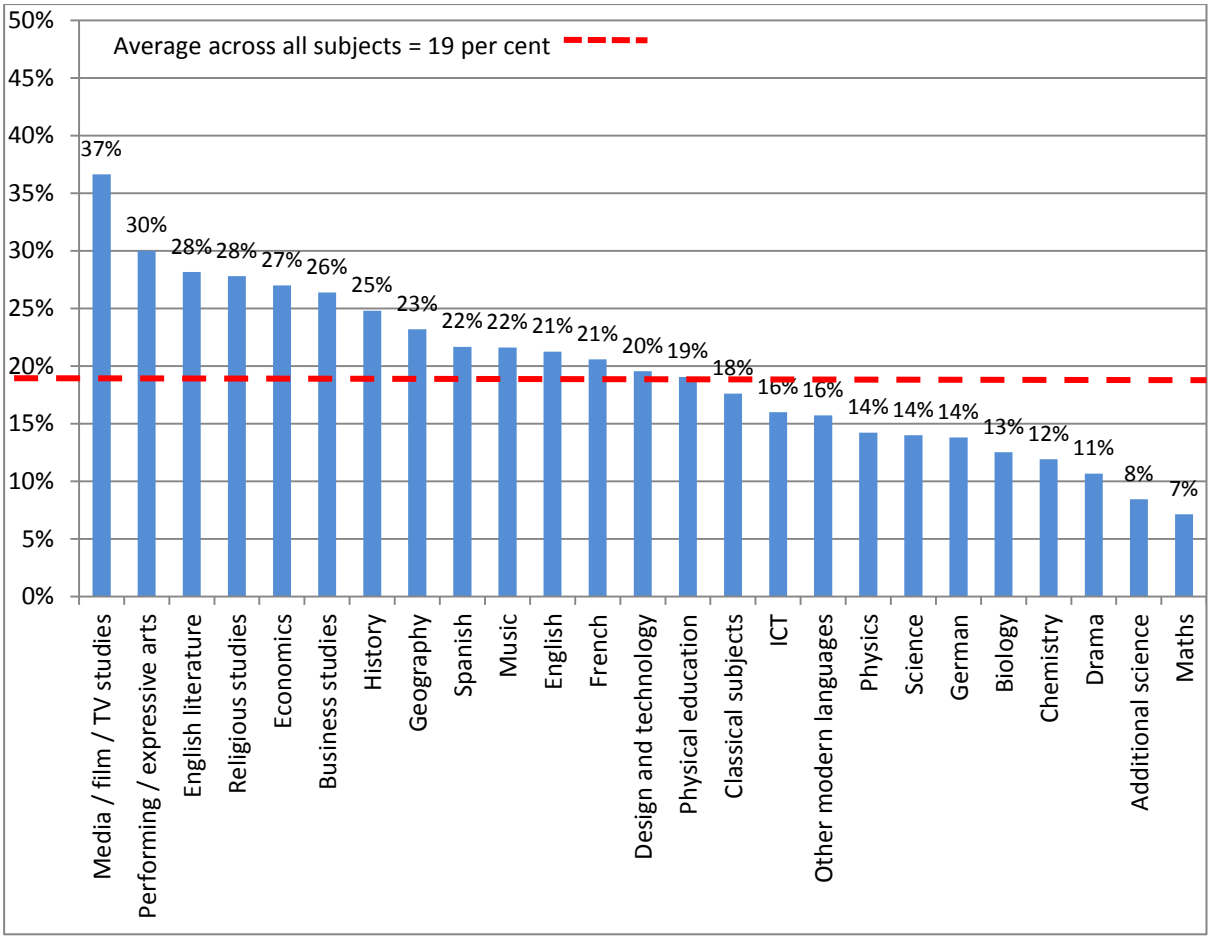
Figure 5: Percentage of A level enquiries about results leading to a unit grade change



Note: Subjects represent Joint Council for Qualifications categories. Data is from summer 2012 for service 2 (priority and non-priority) enquiries about results for all A levels.

There were higher rates of unit grade changes in more subjective subjects. This is likely to reflect difficulties in giving these scripts a definitive mark, rather than marking mistakes. Around 40 per cent of enquiries about results in critical thinking led to a grade change, compared with less than 10 per cent in maths. At GCSE, the rate of unit grade changes appears to be even more closely linked to the inherent subjectivity of the subject. It is notable that some of the subjects with the highest rates of enquiries about results had relatively low grade change rates. These included classical subjects, drama, French and history at A level, and maths at GCSE.

Figure 6: Percentage of GCSE enquiries about results leading to a unit grade change



Note: Subjects represent Joint Council for Qualifications categories. Data is from summer 2012 for service 2 enquiries about results for all GCSEs.

The data above does not give much insight into sources of genuine marking error. Subjective disciplines are more likely to receive unit grade changes than more objective ones. This general trend is supported by other datasets. Appendix A shows the overall reliability coefficients of a number of GCSE and A level units from 2009 to 2011 and 2012. These units usually have reliability coefficients of over 0.8, with a weak trend of higher overall reliability in subjects that are typically more objective, such as science and maths. The trend is also confirmed by examiner performance ratings by subject (see page 45).[27] While high across all subjects, these show a general pattern of higher performance ratings in the more objective subjects.

[27] These ratings are given to examiners by their senior examiner after the marking period. They are based on perceptions of the quality of the examiners' marking and performance on administrative tasks.

4.3.1 Perceived marking errors

Teachers and stakeholders also give us intelligence about where errors are perceived to happen. Some teachers believe there is no one problematic subject, and errors can happen in different subjects or paper types seemingly at random (Oxygen, 2014). However, other teachers believe problems are concentrated in certain subjects, qualifications or parts of the ability range.

Our online survey of teachers conducted in 2013 attracted a very strong response from teachers of modern foreign languages (Dodd, 2014).[28] Their perceptions of GCSE and A level marking were broadly negative, but in their qualitative responses it was clear many were actually more concerned about the grading of these subjects (Dodd, 2014).

This concern over the marking of modern foreign languages is perhaps surprising when we consider some of the data in the system. French and German examiners have higher levels of teaching and examining experience than examiners of any other subject, with around 60 per cent of examiners in both subjects having more than 15 years' examining experience (compared with an average of 22 per cent) (Ofqual, 2013a). According to examiner ratings, French examiners are also among the best performing in the system, with 79 per cent rated one or two (on a scale of one to five, where one is high), compared with an average of 71 per cent. Rates of enquiries about results unit grade changes are also below average in modern foreign languages, particularly at A level.

A high proportion of the teachers who responded to our survey also taught English (Dodd, 2014). English is recognised as one of the more subjective disciplines, particularly at A level, and is, therefore, potentially more difficult to mark reliably. English subjects do have some of the higher rates of unit grade changes as a result of enquiries about results and relatively lower examiner performance ratings (although these are still high).

English teachers and the English Association suggest perceived issues in marking could be linked to the breadth of content assessed in the discipline. While teachers enjoy the academic freedom this offers them, some English literature teachers believe examiners do not always know the full breadth of set texts well enough. In a small, in-depth qualitative study, examiners corroborated that this could occasionally happen, and several cited text unfamiliarity as a challenge to reliable marking (Oxygen, 2014; Dodd, 2014).

[28] Over a third of respondents (35 per cent) taught modern foreign languages, and a further 17 per cent were English teachers.

Teachers, examiners and stakeholder organisations taking part in our research flagged this same issue in other subjects. In particular, some felt multi-topic subjects, such as psychology, history and sociology, were vulnerable to marking errors as a result of examiners being unfamiliar with topics or periods (Oxygen, 2014; Dodd, 2014). All these subjects are subjective to mark, and all have relatively high enquiries about results unit grade change rates at A level. History and sociology have relatively low examiner ratings compared with most other subjects.

From our examiner survey, psychology and sociology examiners had the lowest self-reported levels of teaching and examining experience of any subject, although they could not be described as inexperienced. Eighty-five per cent of both sets of examiners had at least 3 years' examining experience (Ofqual, 2013a). They also consistently reported the lowest levels of confidence in marking and mark schemes. For example, 70 per cent and 72 per cent of sociology and psychology examiners, respectively, agreed with "I receive sufficient briefing about a paper and mark scheme before I begin my marking for each exam*".* This compared with 89 per cent in history.

The data above suggests that, while concerns about marking modern foreign languages may be misplaced, aspects of marking in English, psychology and sociology do show relatively lower levels of examiner performance. Without improved metrics we cannot judge how much this is down to marking mistakes or inconsistencies in marking. However, given the nature of these disciplines, it is likely that much is due to the levels of examiner judgement involved in marking, and the limitations this places on reliability. Nonetheless, text and topic unfamiliarity has been identified as a possible source of error. This could be addressed by better using item-level marking to target questions at examiners with the appropriate expertise.

In terms of qualification type, some teachers feel A level and GCSE marking is poorer than in equivalent qualifications (Dodd, 2014; Headmasters' and Headmistresses' Conference, 2012; Oxygen, 2014). While data is limited, there is no evidence to support this view. Firstly, the profile of examiners across qualification types is very similar. In some cases, the examining experience of examiners marking GCSEs and A levels is slightly below that of examiners of other qualifications, but is still very high. In other areas, such as subject expertise, A levels and GCSEs have among the most qualified examiners.

The responses to our survey of examiners showed examiners of GCSEs and IGCSEs were the most confident about marking and positive about how well they were trained, standardised and supported. In contrast, those marking the IB Diploma were the least confident about their ability to mark and the marking process (although they did have very positive perceptions about their role) (Ofqual, 2013a). There was

also no evidence from our system mapping that marking arrangements in A levels or GCSEs were any less robust than those in equivalent academic qualifications.

Finally, certain teachers expressed concerns about the marking of top-performing students in GCSEs and A levels. They suggested that examiners should be able to recognise a superior knowledge from students, but did not always appear to be able to do so. These perceptions were more likely to be prevalent in independent schools (Oxygen, 2014; Dodd, 2014). It was also a concern among some of the small number of schools and colleges whose cases were heard through our Examinations Procedures Review Service.[29]

This issue of marking top-performing students was raised by stakeholder groups including the Headmasters' and Headmistresses' Conference, Royal Historical Society and English Association. All believed some high-performing students could receive lower marks because they are more likely to give unexpected answers the mark scheme does not capture. This is attributed to the use of prescriptive mark schemes that do not reward some of the higher order skills shown by top-performing students, and to less experienced examiners struggling to mark these responses (Dodd, 2014; Oxygen, 2014).

Examiners did not identify the marking of high-performing students as a particular challenge and they routinely practise marking scripts at the top of the mark range as part of standardisation (Oxygen, 2014; Ofqual, 2013a). They also receive seed scripts of all levels of attainment as part of monitoring on-screen marking. However, there is some very limited evidence in the assessment literature that higher quality responses are more difficult to mark reliably (Pinot de Moira, 2013), as are higher performing students (Pinot de Moira, 2003) and harder items (Sweiry, 2012).

We could find no other evidence in this review to either support or contradict these concerns by teachers. We would need to do a more in-depth study of mark schemes to identify which features support high-quality marking for top-performing students, before we can judge to what extent mark schemes demonstrate these features.

[29] This is the final stage of the post-award appeal process and is a presentation of the case to an appeals panel chaired by senior members of Ofqual, with at least two independent members.

5. Mark schemes

The single most important factor affecting marking reliability is assessment design. At a high level, this relates to the design of whole qualifications: the skills they test and the resulting nature of the tasks within the exams. At a more operational level, this relates to the quality of the design of individual questions and mark schemes. The quality of a mark scheme is central to an examiner's ability to mark well; a poor mark scheme can be the main source of unreliable marking (Meadows and Billington, 2005).

Designing a mark scheme that reflects the full range of student responses, and caters just as well for the highest and lowest achieving students, is no easy task. In extended response tasks that use levels-based mark schemes,[30] capturing every possible response is unfeasible. Examiners must be given clear principles to help them tell apart different levels of student performance. This involves a subjective judgement, which means it is not possible to produce a completely reliable levels-based mark scheme. Despite this, there is evidence that small improvements to the structure, presentation, content and wording of mark schemes could yield some of the biggest improvements in marking reliability.

Various research has been published over the years on what makes a good mark scheme (Boyle *et al.*, 2014; Tisi *et al.*, 2013), but this has been somewhat sporadic and subject-specific. What's more, research has sometimes reported contradictory findings. **Given the importance of mark schemes in securing reliable exams, we will lead a formal programme of research to strengthen the evidence base on what makes a good mark scheme and the features of mark scheme design that have the greatest impact on marking quality for students of all abilities.**

5.1 The quality of mark schemes in general qualifications

A systematic study of the quality of mark schemes was beyond the scope of this research. However, we have considered a range of existing data, which gives an indication of the prevalence and scale of any issues with mark schemes. In 2012, we carried out an indicative investigation into a sample of GCSE and A level assessments in 12 subjects taken in 2011 (see appendix B). This reviewed eight areas affecting the standard of exams. One area related to mark schemes, specifically the ability of mark schemes to produce fair and consistent outcomes.

The study found all assessments to be fit for purpose, but we identified a number of minor issues with assessment design. Over half of these related to mark schemes. Issues included a mismatch in the level of skills and understanding required by

[30] These mark schemes describe a number of levels of response, each of which is associated with a band of one or more marks.

question papers and mark schemes (GCSE history, A level psychology), insufficiently discriminating mark schemes (A level design and technology, A level media studies) and a lack of detail in how to award marks (GCSE psychology, GCSE ICT). This emphasises the importance of quality checking mark schemes and suggests that exam boards could improve this. This information is supported by our 2011 investigation into GCSE and A level question paper errors, which exposed issues with exam boards' assessment design procedures and quality checks (Ofqual, 2011).

Similar issues have been picked up in the accreditation of new specifications in GCSEs and A levels. Recent sample GCSE history assessments showed inconsistencies in all the exam boards' mark schemes. This was most likely to be a mismatch between question papers and mark schemes (for example, mark schemes rewarding higher order skills than identified in the question paper). Less commonly, there were also unclear wording and problems differentiating between students. These issues result in qualifications being refused accreditation.

5.2 Perceptions of mark schemes

Stakeholders' perceptions of mark schemes can give further clues to their effectiveness, or at least indicate how they are viewed by the public. Examiners have a lot of interaction with mark schemes, and exam boards should capture and act on their feedback, both at standardisation (to make minor adjustments to mark schemes in light of the live student responses) and in future assessment design.

Examiners are broadly positive about mark schemes. In our examiner survey, just under nine in ten agreed with "I feel confident when using a mark scheme in my subject (or unit)", although less than three quarters agreed mark schemes were "clear and unambiguous" (Ofqual, 2013a). However, some examiners believed it was sometimes difficult to judge mark schemes in isolation from related standardisation sessions. They explained a more ambiguous mark scheme was clarified by effective standardisation. Mark scheme ambiguity can be a problem when examiners don't feel so effectively standardised (particularly when they are standardised online). Here, the clarity of the mark scheme, and the language used with it, becomes far more important (Oxygen, 2014).

Examiners noted mark schemes could differ quite widely from paper to paper, even within the same exam board or subject type (Oxygen, 2014). Despite this, patterns still emerged. Examiners marking for CCEA or WJEC were significantly more positive about mark schemes than the IB or AQA examiners: 86 per cent of WJEC and 84 per cent of CCEA examiners responded positively to the statement "In my experience, mark schemes are clear and unambiguous." This fell to 65 per cent for the IB examiners and 69 per cent for AQA examiners. Similarly, examiners of maths, chemistry, French, and art and design were the most positive about mark schemes,

in contrast to those marking sociology, psychology and, to a lesser extent, geography (Ofqual, 2013a).

Despite the generally positive perceptions above, mark schemes were identified by examiners as the second biggest challenge to reliable marking (following time pressures). Examiners' comments were inconsistent and did not point to any one issue with mark schemes. Some complained mark schemes were too vague and ambiguous to be able to apply accurately. They called for clearer wording, with clear definitions of terms such as sophisticated and good, used in level descriptors for mark bands. Others suggested that mark schemes were too prescriptive; this was cited, unprompted, by one in ten examiners who responded to our survey of examiners (Ofqual, 2013a). However, neither the quantity nor the strength of responses indicated significant concern about mark schemes among examiners.

Teachers who responded to our online survey were more likely to be critical of mark schemes. Again, comments were inconsistent, with some calling for more constrained mark schemes, and others for more flexible ones (Dodd, 2014). Like examiners, teachers called for clearer wording to describe levels of performance within mark schemes, such as clearly distinguishing between good and excellent knowledge. However, in contrast to examiners, teachers wanted this guidance to better prepare their students for exams (Oxygen, 2014).

In the responses to our call for evidence, some stakeholders raised concerns about mark schemes. The English Association gave examples of mark schemes where there were a small number of bands, poorly worded performance descriptors and inaccurate indicative content. In contrast, the Royal Historical Society was concerned about examiners applying mark schemes inaccurately and inconsistently (Dodd, 2014). Both teachers and stakeholders also expressed concerns about how mark schemes were applied to top-performing students (see section 4).

Finally, information submitted to us by the Independent Schools' Modern Languages Association confirmed there can be fundamental differences in mark scheme design between and within exam boards. Specifically, the Independent Schools' Modern Languages Association identified different rules with applying mark schemes for reading and writing units in French, including capping the marks available for quality of language in relation to marks in other areas. While it acknowledges overly prescriptive mark schemes could invite formulaic responses, the Independent Schools' Modern Languages Association also believes many mark schemes in modern foreign languages are currently too ambiguous. As with some other teachers and examiners, it calls for descriptive words used in mark schemes to be explained better.

The collective intelligence does not point to a significant issue with mark schemes, but the evidence is limited and we will need to carry out further research in this area.

However, it does indicate that the principles behind mark schemes are often inconsistent across subjects and paper types, and mark schemes can vary in quality and are not always subject to rigorous quality checks. It suggests that there is potential for greater checking and consistency of practice with mark schemes. Even relatively small improvements in this area could be significant.

As part of this, we believe exam boards should make better use of item-level data gathered during live marking as well as evidence from the enquiries about results stage to provide feedback on the performance of individual items on an exam paper.

This should be fed into future assessment design more formally and routinely. This feedback loop exists in some exam boards, but it is not universally a formalised process.

Any subsequent last-minute changes to future papers would need to be carefully checked in light of any increased risk of question paper or mark scheme errors. A balance needs to be struck between front-end investment in designing the draft paper and mark schemes, and amendments later on. In any case, there is a real opportunity to strengthen the feedback loop and exploit the ever-increasing data from online systems.

6. Examiners – the people and the role

In April 2013, we surveyed over 10,000 examiners of general qualifications in what is likely to be the biggest ever survey of the workforce. Exam boards told us they used around 51,000 examiners, but we believe the total number of unique examiners working across the seven exam boards at the time was around 34,000. This is based on the fact that 22 per cent of examiners who responded to our survey worked for two or more exam boards (Ofqual, 2013a). This survey, therefore, represents the responses of around one third of the workforce.

Some initial findings from the survey were presented in our first report on marking. In summary, we found the examining workforce to have extremely high levels of subject, teaching and examining experience.

- Examiners had considerable subject expertise. Ninety-two per cent of examiners had a degree or doctorate in the main subject they examined.
- More than 99 per cent of respondents were current or former teachers, many with senior roles: 46 per cent were, or had been, a head of department or above.
- 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.
- Most respondents worked or had worked in comprehensive schools and academies or free schools (54 per cent), with 15 per cent from independent schools.

Examiners' subject, teaching and examining expertise is important for the reliable marking of complex, extended answers, but experience is far less important as questions become less complex (Meadows and Billington, 2007; Tisi *et al.*, 2013). Despite this, UK exam boards are understandably cautious in their use of marking personnel, preferring to use examiners with high levels of expertise almost universally. This is particularly noteworthy given the number of examiners needed.

6.1 Examiner performance

At the end of the marking period, examiners are evaluated on their performance by their senior examiner, both for their quality of marking and their performance on administrative tasks. Most exam boards use a five-point scale, with a rating of one

reflecting the highest level of performance and a rating of three a satisfactory performance.[31]

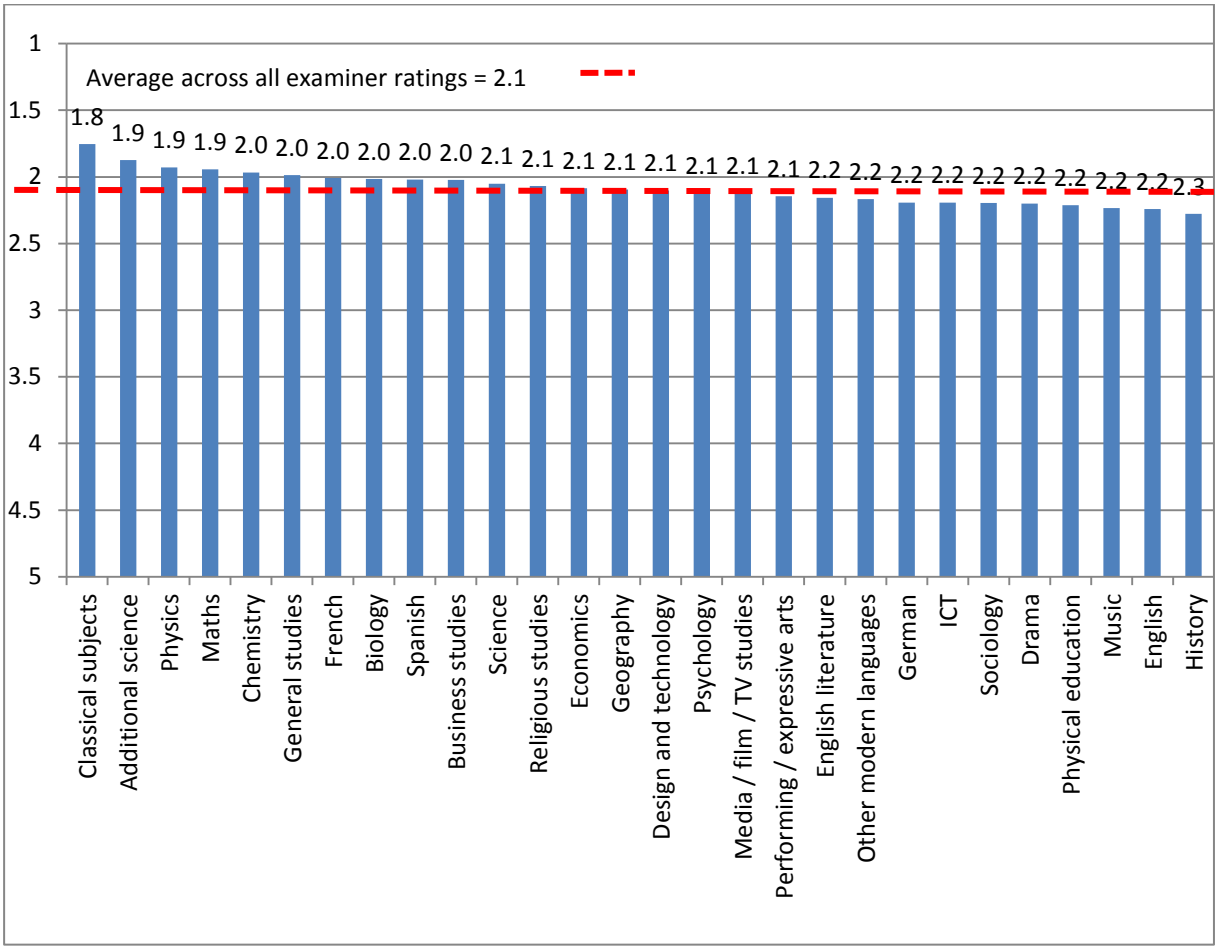
Exam boards use these ratings to decide whether to use examiners again for future exam series. The higher rated the examiner, the more likely they are to be retained. Examiners rated four are unlikely to be used again on the same unit, particularly without retraining. Examiners rated five are not used again. These examiners would usually have been stopped from marking in the marking window and all of their existing work re-marked.

In summer 2012, virtually all examiners (94 per cent) were rated as satisfactory or better, with most (71 per cent) marking to a very high standard and given a rating of 1 or 2. Six per cent of examiners were rated at the lower end of the performance scale (rating of 4 or 5). This pattern was very similar across exam boards and qualifications, although there was some variation by subject.

Figure 7 shows the mean rating of examiners across subjects. The subjects with the lowest rated examiners tended to be subjective disciplines. History examiners had a mean rating of 2.27 and English 2.24. At the other end of the scale, the higher rated examiners marked additional science (1.87), physics (1.92), maths (1.94) and chemistry (1.96). Classical subjects gave an interesting variation to this trend, with the highest mean rating of 1.75.

[31] AQA, CIE, OCR, Pearson Edexcel and WJEC all use a five-point scale to rate examiners. The CCEA uses a four-point scale. The IB evaluates examiners on data collected during the marking period and does not use performance ratings.

Figure 7: Mean rating of examiners by subject, summer 2012



Note: Data is provided for all general qualifications for the exam boards using a five-point scale to rate examiners. It excludes the IB (where no ratings are used) and CCEA (which uses a four-point scale). Subjects reflect Joint Council for Qualifications categories. Only subjects with 300 examiners or more were included in this analysis.

6.2 The role of examiner

Unlike in some other education systems, teachers in England do not have to examine. Individuals decide whether to mark exams, and any associated professional development is not formally recognised. Most examiners are serving teachers. They fit their examining around working hours, usually marking in the evenings, after a day in the classroom, or at weekends.

Managing this examining workload was identified as the most significant challenge for examiners: whether referring to the tight timescales in which to mark a high volume of scripts or the pressures of juggling examining with a teaching role. In our survey, half of examiners felt some or significant pressure to fit examining around other work commitments. One of the most commonly suggested improvements to the system made by examiners was for more time to meet deadlines. Some examiners

wanted to be able to take some time off from their teaching duties to have more time to mark, or make more use of free periods during the school day (Ofqual, 2013a).

Teachers' reasons for becoming an examiner were twofold: wanting to earn additional income, combined with a desire to develop their professional expertise (Ofqual, 2013a). While additional income was the main reason for taking up examining, examiners did not always continue examining for this reason. In terms of developing professional expertise, examiners expressed a desire to learn more about the marking process "and a wish to improve their students' results by better understanding the specification" that they teach (Oxygen, 2014; p. 37). Examiners agreed that gaining this knowledge helped them in their teaching role (Dodd, 2014).

Exam boards find there is currently a good supply of examiners, but in some subjects (usually more subjective disciplines) it is harder to attract and retain suitable examiners. Nonetheless, the supply of examiners is such that all scripts are marked within the timescales available and by a suitably qualified workforce. The upcoming reforms to GCSEs and A levels will bring with them more recruitment challenges. More extended writing in exams may require more expert examiners. In this context, it is crucial the pool of high-quality examiners is as wide as it can be. This means encouraging more teachers to mark general qualifications.

6.2.1 Training and development of examiners

A suitably qualified workforce does not simply rest on the availability of new examiners. It also depends on the development of examiners already in the system. For non-senior examiners, most training focusses on understanding specific papers and mark schemes. This is standardisation. Any other training tends to focus on the technical process of examining, such as how to use online systems. Additional face-to-face training is unusual.

Eighty-eight per cent of examiners who responded to our survey felt examiner training was sufficient. Nonetheless, around a third of examiners agreed that improvements could be made to this training. Some told us there was a need for more and better initial face-to-face training and supplementary guidance materials, particularly for new examiners (Oxygen, 2014; Ofqual, 2013a).

Team leaders receive the same training as examiners. Despite their additional responsibilities, they do not usually receive any formal training in the soft skills needed to manage a team, although this does take place at both Pearson Edexcel and OCR. There can also be a lack of written guidance for team leaders on when and how to contact and feed back to team members. This is perhaps more of a gap in the on-screen marking world, as some examiners told us the move to online standardisation and on-screen marking had put far greater emphasis on the role of

team leader, most notably on their support and feedback during standardisation and the early days of marking (Oxygen, 2014).

Training for senior examiners

The expertise of the most senior examiners (principal examiners, chief examiners and chairs of examiners) is critical to the successful delivery of high-quality question papers and mark schemes. These examiners must have a strong background in assessment as well as the necessary subject and curriculum knowledge. They must have the technical skills to design robust assessments, but also the management skills to lead and train a team of examiners.

From our survey, nine out of ten senior examiners believed they had the skills and training to do the job required of them, and this group told us they were very confident in designing assessments, mark schemes and leading their team (Ofqual, 2013a). Training for senior examiners generally takes the form of mentoring, plus attending ad hoc training events. Only Pearson Edexcel and CIE have formal, structured programmes to train their senior examiners in all aspects of the assessment process. Pearson Edexcel in particular is now requiring its senior examiners to gain a formal qualification in assessment. This is a recent and welcome development in the professionalisation of the workforce. **Whatever forms of training they give, exam boards should continue to make the senior examiner role more professional, making sure senior examiners have the skills and understanding to design high-quality assessments.** This is crucial given the importance of mark schemes in securing reliable assessments, and the varying quality of the mark schemes we have seen.

6.3 Examiner perceptions of the marking of external exams

Throughout all of our research with examiners, it was evident they were a committed, positive and conscientious workforce, who recognised the significance of their work (Oxygen, 2014). Examiners believed the quality of marking of external exams was high. They were highly confident in their own ability to mark accurately and reliably (96 per cent agreed this was the case) and generally believed exams were marked accurately and reliably in their exam board (86 per cent) (Ofqual, 2013a). The most senior examiners were the most confident about marking. They had a better oversight of the process and its various checks and balances, and believed these were effective in delivering accurate and reliable marking.

Many non-senior examiners acknowledged they did not have an oversight of the marking process and could only reflect on their own experience within the system. This was confirmed by in-depth interviews, which showed some examiners did not have a clear picture of the detailed quality controls in the system. Given the important role examiners have in communicating information about marking to teachers and

schools (see section 7), improving examiners' understanding of the technicalities of the whole marking system must be beneficial in improving wider understanding of and confidence in marking (Oxygen, 2014).

Most examiners had a very positive experience of the marking process. Around nine in every ten examiners felt properly supported (94 per cent), had adequate guidance (89 per cent) and felt suitably trained to mark to a high standard (88 per cent). Confidence was higher among examiners marking for certain boards and qualifications. Examiners for CCEA and WJEC consistently gave the most positive responses throughout the survey. They were significantly more positive and confident than examiners marking for the IB. Confidence across qualification types was broadly similar, with the exception of the IB Diploma, where examiners reported the lowest levels of confidence and positivity about the marking process (Ofqual, 2013a).

For all of their confidence and positivity, both senior examiners and examiners had widespread concerns about the increased use of online standardisation. We discuss this in full in section 3.

7. The role of teachers, schools and colleges in marking

Teachers, schools and colleges have a central role to play in the marking system as users of qualifications and teachers as examiners themselves. In this section, we discuss teachers' and other stakeholders' perspectives on marking processes and the role of examiner, as well as their understanding of the marking process. At times, our research highlighted a gap between the reality of the system and how it was perceived by stakeholders. We have, therefore, contextualised our findings with other evidence, where necessary.

7.1 Teachers' perspectives on marking

The majority of teachers are confident in the quality of marking of exams in general qualifications. However, there is a sizeable (and growing) minority who do not share this confidence. In our *Perceptions of A levels, GCSEs and Other Qualifications: Wave 11* survey, 20 per cent of teachers were not confident in the accuracy of A level marking, this rose to 34 per cent for GCSE marking (Ipsos MORI, 2013). These teachers were often extremely critical of the marking system.

Levels of confidence vary across the teaching population. Head teachers are less likely to be confident in marking, as are those teaching in independent schools (Oxygen, 2014; Ipsos MORI, 2013). Perceptions can be poorer among teachers of subjects that are inherently more subjective and, therefore, more difficult to mark reliably, such as A level English literature and history (Oxygen, 2014). That said, the Royal Historical Society stated its "considerable confidence" in the overall reliability of the marking process for general qualifications (Dodd, 2014, p. 30).

As the most common general qualifications, criticisms of marking usually focus on A levels and GCSEs. Teachers in comprehensive schools and academies are often more critical of GCSE marking, while independent schools focus on A level marking (Oxygen, 2014; Dodd, 2014). This may be linked to different pressures on different types of schools and colleges. Teachers in independent and selective schools told us they were under tremendous pressure to help students achieve the highest grades at A level, with a view to securing top university places. Comprehensive schools and academies were more focussed on marking at the C/D grade boundary in core subjects at GCSE due to current accountability measures and the perception that achieving a C grade in certain subjects is important for students' future prospects (Oxygen, 2014).

Most teachers cannot judge the quality of marking of equivalent, less widely taken qualifications, such as IGCSEs, the Pre-U and the IB Diploma. Where they do give a view, confidence in the marking of these qualifications is higher than for GCSEs or A levels (Dodd, 2014). Some of the teachers we interviewed from independent schools

had moved to some equivalent qualifications in the hope this would deliver better quality of marking, particularly for top-performing students (Oxygen, 2014). As discussed in section 4, we found no evidence in this review that the marking of equivalent exams was any more robust than the marking of A levels and GCSEs. This appears to be a misconception among teachers.

7.1.1 What affects teachers' perceptions of marking?

If students' final grades broadly match predicted or forecast grades[32] and are in line with those achieved by similar cohorts, teachers' perceptions of marking are positive. Where this is not the case, teachers often blame marking. However, this view that unexpected results must be caused by poor marking is a misconception. Even assuming predicted grades are accurate and unbiased (in fact over 40 per cent of all predictions are over-predicted)[33], there are many factors that influence students' final grades aside from marking, not least how the students perform. In an assessment system with an emphasis on valid, stretching assessments, it is not possible to control all these factors to deliver perfectly reliable results. Nonetheless, for many, the predictability of student results and quality of marking of exams are one and the same issue.

Most teachers in our sample believed student grades were in line with predictions in 95 per cent of cases. Most teachers saw their perceived level of error (around 5 per cent) as almost inevitable in an exam system of this scale, particularly in more subjective disciplines. They expected blips in marking from time to time. This perceived level of error is actually higher than enquiries about results data suggests (Ofqual, 2013d). These teachers felt the current level of perceived marking accuracy was acceptable. Some recognised that marking could be made more reliable through changing the nature of exams, but noted this would lead to a form of education and assessment they would not support (Oxygen, 2014).

We should acknowledge that this perceived measure is very different from a true level of marking accuracy. Enquiries about results data (which is likely to significantly over-estimate marking mistakes) shows that, in summer 2013, a little over 1 in 200 A level and GCSE scripts received a qualification grade change as a result of an enquiry about results.

---

[32] Predicted grades are the grades provided by schools and colleges to UCAS for the purposes of university entry, forecast grades are provided by schools and colleges to exam boards.

[33] A study by the Department for Business, Innovation & Skills found 52 per cent of predicted A level grades were accurate in 2009. Just under 90 per cent of grades were accurately predicted to within one grade. Forty-two per cent of all predictions were over-predicted by at least one grade (Department for Business, Innovation & Skills, 2011). This was supported by a study by Gill and Chang (2013), which found that in summer 2012, 48 per cent of forecast grades provided to OCR were accurate, with 39 per cent over-predicted by at least one grade.

Teachers' confidence in marking can also be undermined by experience of the enquiries about results process. Most significantly, confidence in marking is damaged by the experience of large mark changes resulting from an enquiry about results. These mark changes are rare. In summer 2012, less than 1 per cent of mark changes made as a result of an enquiry about results in French and geography A levels (and equivalent exams) were changes of 10 per cent or more of the total raw mark available for the paper. But, these big mark changes, understandably, have a big impact on confidence. Similarly, experience of submitting an enquiry about results and receiving no mark change (when a teacher believes a change is justified) can also undermine confidence in the wider marking system.

To a lesser extent, confidence is also undermined by common misconceptions about marking. Most notably, this includes myths about examiners with no subject and teaching experience (Oxygen, 2014).

7.1.2 Knowledge and understanding of marking

Understanding of the marking process among teachers is mixed. Some teachers have examining experience and a clearer understanding of how marking works. Those without this experience are far less likely to understand the marking process. In our recent survey of teachers, self-reported understanding of marking varied considerably. Most teachers (59 per cent) rated their knowledge of the marking process at seven out of ten, or higher. However, almost 30 per cent of our respondents put the figure at four out of ten, or lower. For some teachers, there also seemed to be a mismatch between self-reported and observed knowledge of marking, and we could see clear misunderstandings about marking. It was also clear some teachers had difficulty with the differences between marking and grading (Dodd, 2014).

In our in-depth interviews with teachers, most acknowledged they had little or no understanding of how external marking worked in terms of detailed quality control processes. The often outdated impressions of many had come from colleagues who were marking or had marked in the past. Many teachers had no knowledge of on-screen or item-level marking. Furthermore, this research found teachers were not always interested in learning about the mechanics of the marking process.

Generally, the information they sought from exam boards was to help them better prepare students for exams. But when details of marking quality controls were explained to them, this greatly improved perceptions of the sophistication of the system. Where an exam board strongly promoted its checking systems to teachers, or where a school or college had a number of teachers who were examiners and knew the system, confidence tended to be higher. The more teachers knew about the working of the system, the more likely they were to trust it (Oxygen, 2014).

The picture above does present a tension. As end users of qualifications, teachers can give a valuable insight into how qualifications are performing, and where improvements can be made. However, the evidence above suggests that, for many, there is a gap between perceptions and the reality of the system. In such cases, the most critical improvement we can identify is to improve the knowledge and understanding of marking and assessment in the teaching profession more widely. **It is important teachers trust the system and are informed and engaged users of it. The exam system is not just the responsibility of the exam boards and qualifications regulator. Schools and teachers should play their part in marking by actively learning more about the marking system and supporting and participating in examining work.**

To improve understanding of the system, we must first distinguish between areas where teachers have established misconceptions about the marking system, and those where they simply have a lack of knowledge and information. This can be difficult to find out, particularly where levels of understanding vary so widely.

Perhaps the most obvious misconceptions centre on examiners. As discussed previously, teachers believe examiners should be experienced, current teachers and subject experts. Despite the extremely high levels of experience in our examining workforce, many do not believe this is the case. Our 2013 survey of teachers found just 54 per cent of respondents believed examiners were subject experts (Dodd, 2014). This is in stark contrast to the evidence presented in section 6, which shows examiners have extremely high levels of subject, teaching and examining experience. When proposing improvements to the marking system, teachers are most likely to refer to improvements in marking personnel. They are clear that examiners should be experienced, serving teachers who are specialists in their subject. Not all believe this is the case. This is a fundamental misconception, which undermines confidence in marking.

In terms of more specific aspects of the marking process, teachers hold some less significant misconceptions about different types of marking. Many are not aware of the prevalence of on-screen, item-level marking, and tend to imagine marking as a traditional pen-and-paper exercise. This aside, few teachers seem to have preconceived ideas about the technicalities of the system, unless they have examining experience. Their knowledge and understanding here are low, and they have little sense of how marking is managed or quality assured. In our survey of teachers, only 44 per cent thought marking was monitored throughout the marking period compared with 43 per cent that thought it was not. A further 40 per cent did not agree that examiners were trained in how to apply mark schemes correctly (Dodd, 2014). This appears to be a gap in information and understanding, rather than a misconception.

Exam boards should circulate clear, up-to-date information on marking processes and quality controls to teachers, schools and examiners. They can also make better use of their examiners to pass on information within schools and colleges.

7.1.3 Improving the system

Given that many teachers have a limited insight into the detailed workings of the marking system, few were able to suggest tangible improvements to the process. Where improvements were proposed, they were most likely to refer to examiners: their training, pay and experience. Teachers also emphasised that exam boards must identify and remove rogue examiners (Oxygen, 2014). Around one in ten teachers also proposed improvements to mark schemes (Dodd, 2014).

As discussed in section 4, teachers in specific subjects, including English, psychology and sociology, also pointed to problems with the breadth of subject knowledge needed by examiners. Their suggestions were to set fewer topics and give marking of texts and topics only to examiners who knew them well (Oxygen, 2014).

The Association of School and College Leaders made the case for improving the quality of marking through giving assessment “a higher profile within the professional framework of the teaching profession”, and through giving greater support and recognition to their examiners. They also believed there should be more consistency between exam boards in how they recruit, train and performance manage senior examiners (Dodd, 2014).

In May 2013, the Headmasters' and Headmistresses' Conference gave us five detailed recommendations to improve the marking and assessment system. These are discussed in full in our formal response to the Headmasters' and Headmistresses' Conference, issued in February 2014.[34]

7.2 Teachers’ attitudes to examining

The quality of the marking system rests, in part, on the availability of suitably qualified examiners. It is crucial that enough teachers are willing to mark general qualifications, and their support of the exam system is backed by schools and colleges. However, in-depth interviews found, in our sample of teachers at least, teachers do not always hold a positive view of marking. On the whole, examining has a reputation as:

[34] www.ofqual.gov.uk/documents/our-letter-to-hmc-about-quality-of-marking

> an unenjoyable slog that affects personal life negatively. Teachers who don't examine can be afraid to take it on… The experience of marking is imagined to be stressful and unpleasant (Oxygen, 2014, p44).

This impression came from feedback from other teachers who examined or had examined and was backed up by historical difficulties experienced by exam boards when recruiting in certain subjects. This has led some teachers to believe "nobody wants to examine" (Oxygen, 2014, p44).

Teachers in our discussion groups had polarised views about examining. Some saw examining as a leg-up for their teaching career, and some reported their senior teachers who were examiners had guru status in their school. However, others perceived examining as a lower status activity, "good for NQTs [newly qualified teachers] for a year or so" and useful for more junior teachers earning lower salaries (Oxygen, 2014, p. 44).

Some of these views of examining are unfortunate. While marking does have its challenges, many current examiners told us how much they enjoyed examining, valued the importance of their work and noticed the benefits it had on their teaching: "I enjoy the challenge very much. [It is] a great privilege to contribute" (Ofqual, 2013a).

7.3 Schools and colleges' attitudes to examining

The likelihood of teachers examining is often driven by the attitude of their school or college, or head teacher. Examiners rarely decide to start examining based on the encouragement of their school or college. Just 9 per cent of examiners cited this as a reason for taking up examining (Ofqual, 2013a). But a lack of support can be critical in deterring teachers from marking.

Teachers, examiners and other stakeholders told us schools and colleges' attitudes to examining varied. Almost a fifth of examiners (17 per cent) did not believe their school or college gave them enough support for examining work. Newer examiners felt this need for support more keenly, particularly as they were more likely to struggle to manage marking alongside work commitments. Across our research, selective and independent schools appeared to be more supportive of examining than comprehensives and academies. Eighty-seven per cent of examiners from independent schools felt supported by their school, compared with comprehensives (81 per cent) and academies and free schools (79 per cent) (Ofqual, 2013a).

Many schools and head teachers were aware of the benefits examining can bring. These schools recognised that marking gives their staff a deeper understanding of a specification and the way in which marks are awarded. Some referred to this as insider information (Oxygen, 2014, p. 45). Teachers in selective state and independent schools appeared to be more likely to use examining strategically to

deliver improved results in their school. In some cases, this included encouraging staff from every department to examine.

Although most schools and colleges support examining, we hear others are indifferent or even hostile towards it. Some teachers told us their school actively discouraged examining. These negative attitudes appear to be driven by concern that examining will reduce the time teachers have available for teaching and other responsibilities at their school or college. This can result in conflict between examiners and their school or college. Examiners might be refused permission to attend standardisation and, in the most extreme cases, teachers may even be disciplined for examining (Oxygen, 2014; Ofqual, 2013a). These attitudes to examining are concerning. All schools and colleges rely on the exam system, and they must be willing to support this system if it is to be as good as it can be.

7.4 Attracting more teachers to become examiners

All the teachers interviewed agreed examining was the best route to understanding how marks are awarded in external exams, and, therefore, an excellent way for teachers to learn how to improve their students' preparation for exams. They suggested exam boards should emphasise these benefits in their recruitment campaigns. Teachers also recommended exam boards did more direct outreach activities with schools and colleges, using existing forums such as local teachers' union groups. Some head teachers said they would value a more flexible approach to reimbursing schools for the cost of providing cover when an examiner needed to attend meetings during term time (Oxygen, 2014).

Examiners in our interviews and discussion groups believed their rate of pay was not particularly good, although marking the most straightforward types of papers was sometimes seen as better value. Teachers and some stakeholder groups also suggested that the remuneration offered and the time pressures involved made examining relatively unattractive for serving teachers (Dodd, 2014). Pay is undoubtedly a factor when considering becoming an examiner. However, it is also clear that teachers become, and remain, examiners for a wide range of reasons, including professional development, social interaction and improved results for their students (Oxygen, 2014). Therefore, increasing pay rates by a marginal amount seems unlikely to attract more teachers to become examiners. Raising pay rates by a bigger amount would have real implications for the cost of general qualifications to schools and colleges.

Therefore, improved marketing and outreach from exam boards on the benefits and profile of examining and greater support from schools and colleges to manage the dual teaching and examining roles are likely to be more effective and sustainable in securing an experienced and committed pool of examiners.

8. The enquiries about results and appeals system

No matter how strong the marking system, there will always be a need for an appeals mechanism. We know marking can never be completely free from error. In any complex system of this scale, it is inevitable that mistakes will be made, beyond the inevitable differences of judgement between different examiners. To complement the arrangements for checking marking, as set out previously, we need robust systems to let schools and colleges challenge possible marking mistakes. In the case of GCSEs and A levels, this is the enquiries about results and appeals system. Equivalent qualifications, including the IB Diploma, Pre-U Diploma and IGCSEs, have similar arrangements in place at the enquiries about results stage.

If it wishes to challenge a mark, a school or college can submit an enquiry about results, followed up, if needed, by up to three appeal stages. In summer 2013, schools and colleges in England, Wales and Northern Ireland submitted 301,300 enquiries about results for external assessments in GCSEs and A levels, related to 2.3 per cent of all scripts. Of these, 16.5 per cent of qualification results involved in enquiries about results led to a qualification grade change. This represented 0.6 per cent of all certifications (Ofqual, 2013d). If a school or college is not satisfied with the outcome of an enquiry about results, it can submit an appeal. There were 493 stage 1 appeal cases in summer 2012,[35] of which 41 progressed to stage 2. Of these, eight cases progressed to the Examinations Procedures Review Service, which we run.

The enquiries about results and appeals system is complex. Altogether, there are seven services or appeal stages for schools or colleges to pursue. These are set out in figure 8. Requirements for processing enquiries about results and appeals are set out in our *GCSE, GCE, Principal Learning and Project Code of Practice*.[36]

[35] One appeal case may be for one or more students, so the number of students involved will be higher than the 493 cases.

[36] See www.ofqual.gov.uk/files/2011-05-27-code-of-practice-2011.pdf . Please note the section of the *Code of Practice* dealing with enquiries about results and appeals does not apply to IGCSEs, the Pre-U Diploma, International A levels and the IB Diploma. However, all providers of these qualifications must comply with the *General Conditions of Recognition*.

Figure 8: The process for handling challenges to marking through enquiries about results and appeals

Service or appeal type	Detail of the process
Enquiries about results: service 1 In summer 2013, 1,950 enquiries about results were made at unit level (A level and GCSE).	This is a clerical check to make sure each question has been marked and all the marks have been totalled correctly.
Enquiries about results: service 2 In summer 2013, 277,200 service 2 enquiries about results (for A level and GCSE) and 22,100 priority service 2 enquiries about results (for A level) were made at unit level.	This is a review of the original marking undertaken by an examiner who is usually more senior than the original one. It includes a clerical check. The reviewing examiner can see the marks and annotations of the first examiner and judges if the marks have been awarded correctly. At A level, reviews can be fast-tracked as a priority service 2 where a place at a higher education institution is at stake.
Enquiries about results: service 3 In summer 2013, 3,100 enquiries about results were made at unit level (A level and GCSE).	This is a review of an exam board's moderation of internal assessment, to make sure any adjustments made by the moderator were fair and appropriate.
Appeal – stage 1 In summer 2012 (the latest year for which figures on appeals are available), 493 stage 1 appeals were submitted (A level and GCSE).	A preliminary review by a senior member of the exam board who has not been involved with the case. This considers if the exam board's processes are consistent with the *GCSE, GCE, Principal Learning and Project Code of Practice* and the exam board's own published procedures.
Appeal – stage 2 In summer 2012, 41 appeals progressed to stage 2 (A level and GCSE).	A presentation of the case by the school or college to an appeals panel convened by the exam board, with at least one independent member.
Examinations Procedures Review Service In 2013, the Examinations Procedures Review Service accepted eight cases for a hearing related to the 2012 summer series (A level and GCSE). Two were upheld.	This is the final stage of the appeals process and is a presentation by the school or college (or private student) of the case to an appeals panel chaired by senior members of Ofqual, with at least two independent members.

8.1 Criticisms of the enquiries about results and appeals system

The enquiries about results and appeals system has been in place for a number of years, although it has been adjusted over time. Recently, it has been criticised, particularly where the number of marks challenged has risen substantially (Ofqual, 2013d). Some head teachers and teachers are concerned about the philosophy, fairness and workings of the system, and lack confidence in the outcomes. We share some of their concerns. In this review of marking, we do not attempt to cover all criticisms of the system, nor diagnose their causes, but some of the main issues are summarised below.

One of the many concerns of teachers is the complexity of the system. There are potentially four stages for challenging a mark: firstly through the enquiries about results process, then through two appeal stages, and finally by applying to the Examinations Procedures Review Service. At each stage, the focus of the appeal is slightly different. While the enquiries about results process checks for errors in the application of the mark scheme, subsequent appeal phases focus on whether processes have been followed correctly and the *GCSE, GCE, Principal Learning and Project Code of Practice* adhered to.

Perhaps linked to this complexity, some see the system as opaque, with perceptions of exam boards hiding behind the process at both the enquiries about results and appeals stages (Headmasters' and Headmistresses' Conference, 2012), and being reluctant to initiate a whole class/cohort extended review.[37] Some teachers also complain about the cost (there is a fee if a challenge is not successful) and timeliness of the process. Many enquiries about results are completed fairly swiftly and almost always within the specified time frame (Ofqual, 2013d). Later appeal stages can go on much longer, although delays are not always caused by exam boards.

One major criticism of the enquiries about results system specifically relates to the very philosophy behind the process. At present, exam boards are instructed to carry out a review of the original marking to make sure the agreed mark scheme has been applied correctly. This is not the kind of blind re-mark that many schools and colleges would like to see. The Headmasters' and Headmistresses' Conference, in particular, calls for the *GCSE, GCE, Principal Learning and Project Code of Practice* to be revised to require exam boards to carry out an independent external blind re-mark to avoid any claims of this system being a “mere rubber stamping exercise by the exam board” (Headmasters' and Headmistresses' Conference, 2012, p. 7).

[37] Joint Council for Qualifications guidelines state the exam board will authorise an extended review if a trend of significant under-marking is revealed. This is generally defined as a change of more than 5 per cent of the raw marks for the paper. At least 50 per cent of the school or college’s sample must have experienced significant under-marking for it to be considered a trend (JCQ, undated). Anecdotally, some exam boards state they often do not wait for these thresholds to be reached before authorising an extended review.

Certainly, the current approach to reviewing marking is at odds with what happens in the live marking window. The majority of monitoring of live marking takes place through a modified form of blind re-marking. The reviewing examiner cannot see the marks and annotations of the first examiner, and marks the work uninfluenced by the thinking of the first examiner. Research shows this is likely to give a better measure of marking inconsistency or inaccuracy (Billington, 2012). Furthermore, the review of marking at enquiries about results may have other unintended consequences. Anecdotally, exam boards believe examiners may look for extra marks to award students, despite being briefed not to do so. Examiners are aware of the weight riding on student grades, and that many students for whom an enquiry about results is submitted are likely to be just below a grade boundary.

8.2 Inconsistencies in the enquiries about results system

As well as the common criticisms above, our own analysis of the enquiries about results system, in particular, has highlighted some further inconsistencies and unhelpful practices.

As we have discussed, in a valid assessment system it is possible to have legitimate differences in opinion on the mark to be awarded between two qualified and skilled examiners. This is recognised in the live marking period by most exam boards by applying a unit-specific marking tolerance. In the enquiries about results process, no such formal numerical tolerance is applied (with the exception of CIE). Examiners are instead asked to review the script to make sure the mark scheme has been applied correctly. We have seen that around four fifths of mark changes made as a result of an enquiry about results in summer 2012 fell within the original marking tolerance (section 4). While some of these enquiries about results will contain small marking mistakes, this is significant as it suggests that many enquiries about results mark changes are likely to represent a legitimate difference in opinion between examiners.

Many schools and colleges are aware of this anomaly. Our interviews with teachers found some admit to using the enquiries about results system speculatively where students' marks fall just below an important grade boundary. Comprehensive schools and academies are most likely to challenge marks just below the crucial C/D grade boundary in core GCSE subjects. Selective independent and state-maintained schools choose to focus enquiries about results at students who are just below the A/B or A*/A grade boundary with a view to securing top university places (Oxygen, 2014). This is supported by the enquiries about results data for the summer exam series 2013, in figures 9 and 10. It shows, at GCSE, by far the greatest number of enquiries about results were received for students at grade D, whereas at A level most enquiries about results were received at grade B (Ofqual, 2013d).

Figure 9: Percentage of qualification grades involved in enquiries about results for GCSE, summer exam series 2013

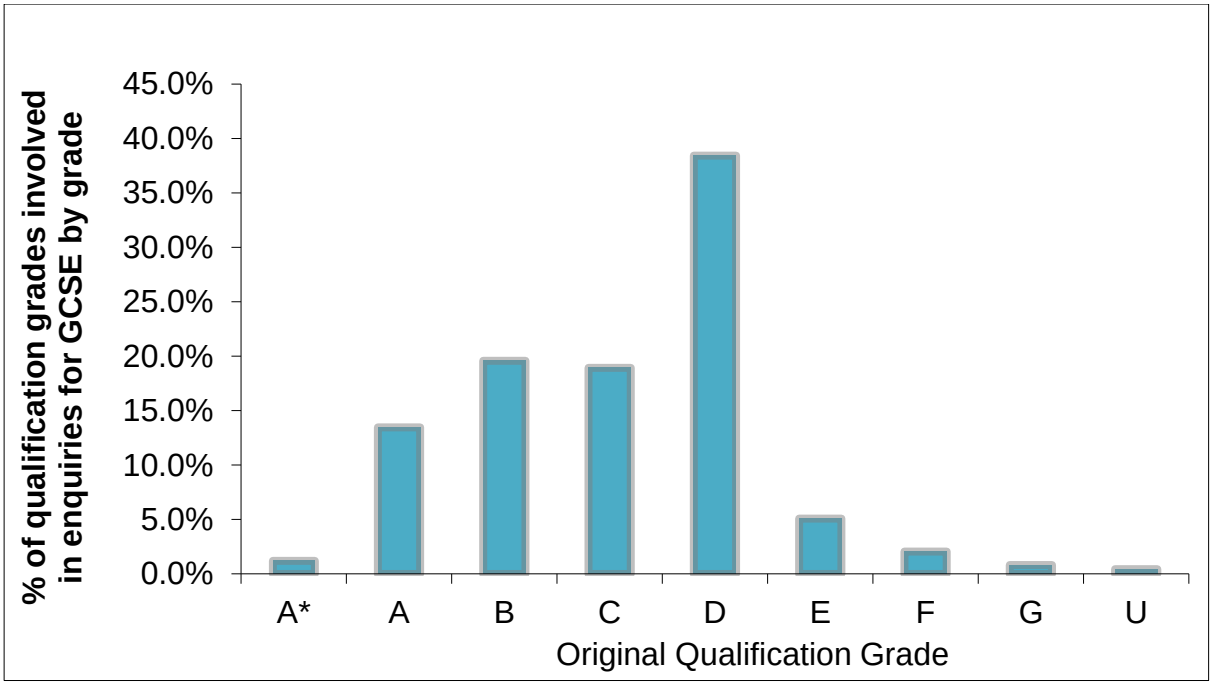
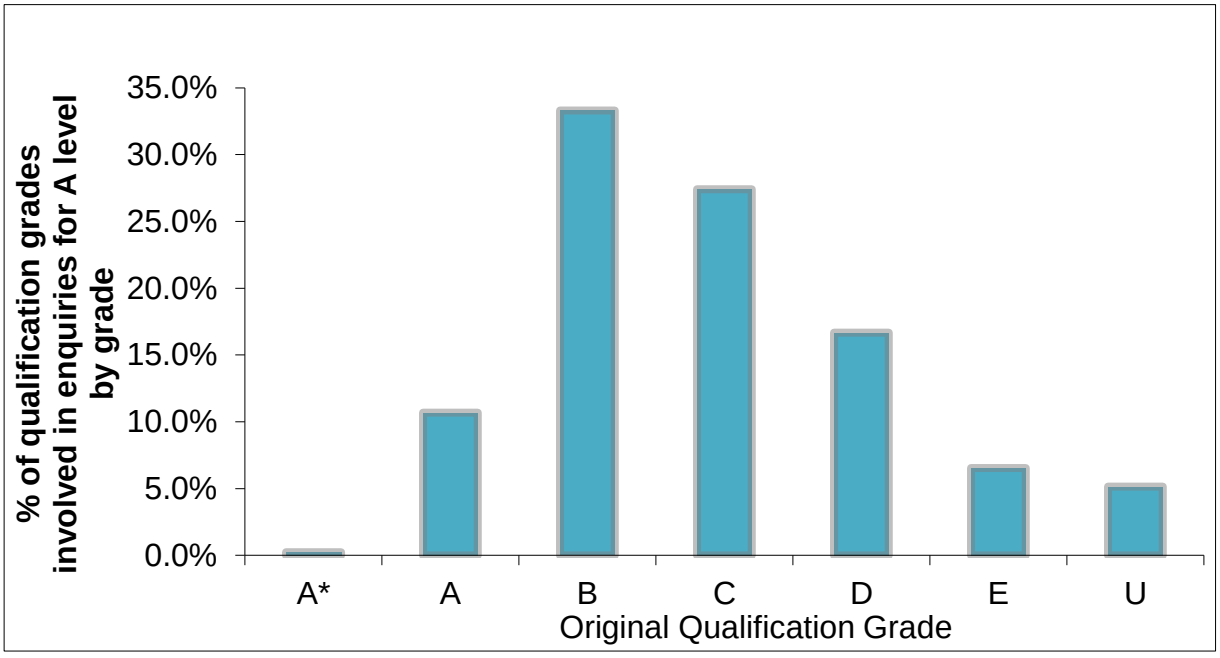


Figure 10: Percentage of qualification grades involved in enquiries about results for A level, summer exam series 2013



Schools and colleges that are aware of the subjectivity in marking know an enquiry about results may lead to a change in marks. Where students' marks are just below a key grade boundary, the likelihood is their grade will improve or, at worst, stay unchanged as a result of the enquiry. This led some head teachers to describe the

practice of entering enquiries about results just below grade boundaries as a "one way bet" (Oxygen, 2014, p. 12). As the pressure on schools and colleges to deliver results increases, so do the incentives to appeal marking.

Therefore, a high (and potentially increasing) volume of enquiries about results are, we believe, motivated by a speculative attempt to improve results, for whatever reason. This is not what a system of redress is intended for. While the enquiries about results system is currently coping with this increasing volume of enquiries, there is a risk it could eventually struggle under the pressure of this misuse.

Given the significant criticisms above, it is clear the current enquiries about results and appeals system is in need of review. Processes are coming under increasing pressure, and we will make changes so they are fit for the future. We will carry out a fundamental re-design of the system and consult on changes to our formal regulatory requirements. We will aim for this new system to be in place for the summer 2015 exam series. Such a system must be timely, fair and transparent, and robust enough to tell apart legitimate variations in marks and marking mistakes. It also needs to recognise the pressures and incentives on schools and colleges and, as far as possible, avoid encouraging speculative enquiries.

9. References

Baird, J., Greatorex, J. and Bell, J. F. (2004) *What Makes Marking Reliable? Experiments with UK Examinations, Assessment in Education: Principles, Policy & Practice,* 11, 3, 331 to 348. University of Bristol, Bristol. Available at: http://research-information.bristol.ac.uk/en/publications/what-makes-marking-reliable-experiments-with-uk-examinations(332a3f96-53e6-4062-80ed-eef64f01187c)/export.html (accessed 6th February 2014).

Baird, J., Black, P., Bèguin, A., Pollitt, A. and Stanley, G. (2012) *The Reliability Programme: Final Report of the Technical Advisory Group.* Coventry, Ofqual. In D. Opposs and Q. He (eds.) *Ofqual's Reliability Compendium,* pp. 771 to 838. Coventry, Ofqual. Available at: www.ofqual.gov.uk/standards/research/reliability/compendium (accessed 6th February 2014).

Billington, L. (2012) *Exploring Second Phase Samples: What Is the Most Appropriate Basis for Examiner Adjustments?* Manchester, AQA, Centre for Education Research and Policy. Available at: https://cerp.aqa.org.uk/sites/default/files/pdf_upload/CERP-RP-LB-01062009.pdf (accessed 6th February 2014).

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

Boyle, A., Alton, A., Mitchell, T., Murphy, R. and Dalby, D. (2014) *Standardisation Methods, Mark Schemes, and Their Impact on Marking Reliability*. Coventry, Alpha*Plus* Consultancy Ltd for Ofqual. Available at: www.ofqual.gov.uk/documents/standardisation-methods-mark-schemes-and-their-impact-on-marking-reliability (accessed 12th February 2014).

Bramley, T. and Dhawan, V. (2012) *Estimates of Reliability of Qualifications*. Coventry, Ofqual. In D. Opposs and Q. He (eds.) *Ofqual's Reliability Compendium*, pp. 523 to 556. Coventry, Ofqual. Available at: www.ofqual.gov.uk/standards/research/reliability/compendium (accessed 6th February 2014).

Brooks, V. (2004) *Double Marking Revisited*, *British Journal of Educational Studies*, volume 52, issue 1, pp. 29 to 46. [no place], Wiley Online Library. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8527.2004.00253.x/abstract (accessed 6th February 2014).

Center for Educator Compensation Reform (2012) *Measuring and Promoting Inter-rater Agreement of Teacher and Principal Performance Ratings*. Manchester, AQA, Centre for Education Research and Policy. Available at: http://cecr.ed.gov/pdfs/Inter_Rater.pdf (accessed 6th February 2014).

Chamberlain, S. and Taylor, R. (2010) *Online or Face-to-face? An Experimental Study of Examiner Training, British Journal of Educational Technology,* 42(4), pp. 665 to 675. [no place], Wiley Online Library. Available at: http://onlinelibrary.wiley.com/doi/10.1111/bjet.2011.42.issue-4/issuetoc (accessed 6th February 2014).

Dhawan, V. & T. Bramley (2013). *Estimation of inter-rater reliability*. Office of Qualifications and Examinations Regulation: Coventry. Available at: www.ofqual.gov.uk/files/2013-01-17-ca-estimation-of-inter-rater-reliability-report.pdf. (accessed 10th February 2014).

Department for Business, Innovation & Skills (2011) *BIS Research Paper Number 37 Investigating the Accuracy of Predicted A Level Grades as part of 2009 UCAS Admission Process.* [no place], Department for Business, Innovation & Skills. Available at: www.gov.uk/government/uploads/system/uploads/attachment_data/file/32412/11-1043-investigating-accuracy-predicted-a-level-grades.pdf (accessed 6th February 2014).

Dodd, L. (2014) *Quality of Marking in General Qualifications - Survey of Teachers 2013.* Coventry, Ofqual. Available at: www.ofqual.gov.uk/documents/quality-of-marking-in-general-qualifications-survey-of-teachers-2013 (accessed 13th February 2014).

Evers, A., Sijtsma, K., Lucassen, W. and Meijer, R. (2010) *The Dutch Review Process for Evaluating the Quality of Psychological Tests: History, Procedure, and Results, International Journal of Testing* 10, pp. 295 to 317. [no place], Taylor & Francis Online. Available at: www.tandfonline.com/doi/abs/10.1080/15305058.2010.518325 (accessed 13th February 2014).

Feldt, L. and Brennan, R. (1989). Reliability. In *Educational Measurement* (3rd Edition, edited by R. Linn), pp. 105-146. The American Council on Education, MacMillan.

Frisbie, D. A. (1988) *Reliability of Scores from Teacher-made Tests, Educational Measurement: Issues and Practice,* 7(1), pp. 25 to 35. [no place], National Council on Measurement in Education. Available at: http://ncme.org/linkservid/65BD2D34-1320-5CAE-6E13B6D9BD1AB46A/showMeta/0/ (accessed 6th February 2014).

Gill, T. and Chang, Y. (2013) *The Accuracy of Forecast Grades for OCR A levels in June 2012*, *Statistics Report Series No.64*. Cambridge, Cambridge Assessment. Available at: www.cambridgeassessment.org.uk/Images/150215-the-accuracy-of-forecast-grades-for-ocr-a-levels-in-june-2012.pdf (accessed 7th February 2014).

Haladyna, T. M., Rodriguez, M. C. and Downing, S. M. (2013) *Developing and Validating Test Items*. New York, Routledge.

Hayes, M. and Pritchard, J. (2013) *Estimation of Internal Reliability.* Coventry, Ofqual. Available at: www.ofqual.gov.uk/documents/estimation-of-internal-reliability/all-versions (accessed 6th February 2014).

He, Q., Hayes, M. and Wiliam, D. (2011). Classification accuracy in Key Stage 2 National Curriculum tests in England. *Research Papers in Education* 28, 22–42. Ofqual, Coventry. Available at http://webarchive.nationalarchives.gov.uk/+/http://www.ofqual.gov.uk/files/reliability/11-03-16-Ofqual-Classification-Accuracy-in-Results-from-Key-Stage-2-National-Curriculum-Tests.pdf (accessed 10th February 2014).

Headmasters' and Headmistresses' Conference (2012) *England's 'Examinations Industry': Deterioration and Decay: A Report from HMC on Endemic Problems with Marking, Awarding, Re-marks and Appeals at GCSE and A level, 2007-12.* [no place], HMC. Available at: www.hmc.org.uk/?s=deterioration+and+decay (accessed 6th February 2014).

Hutchison, D. and Benton, T. (2012). Parallel universe and parallel measures: estimating the reliability of test results. Cited in *Ofqual's Reliability Compendium* (D. Opposs and Q. He), pp. 419-458. Ofqual, Coventry. Available at www.ofqual.gov.uk/standards/research/reliability/compendium/ (accessed 10th February 2014).

Ipsos MORI (2013) *Perceptions of A levels, GCSEs and Other Qualifications: Wave 11*. Coventry, Ofqual. Available at: www.ofqual.gov.uk/standards/statistics/perceptions (accessed 6th February 2014).

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*, 18, 2, pp. 107 to 124. Cambridge, Cambridge Assessment. Available at: http://editlib.org/p/110748/ (accessed 6th February 2014).

Johnson, M., Nádas, R. and Bell, J/ (2010) Marking essays on screen: An investigation into the reliability of marking extended subjective texts. *British Journal of Educational Technology* 41, 814–826.

Johnson, M., Hopkin, R., Shiell, H. and Bell, F. (2012) Extended essay marking on screen: is examiner marking accuracy influenced by marking mode? *Educational Research and Evaluation* 18, 107-124

Joint Council for Qualifications (undated) *GCSE, GCE, Principal Learning & Projects (including Extended Project). Post-Results Services. Information and guidance to*

centres for examinations taken in: *June 2013, November 2013 and January 2014.* JCQ. Available at: http://www.jcq.org.uk/exams-office/post-results-services (accessed 10th February 2014).

Knoch, U., Read, J. and von Randow, J. (2007) *Re-training Writing Raters Online: How Does It Compare with Face-to-face Training? Assessing Writing*, 12(1), pp. 26 to 43. [no place], [no publisher]. Cited by: Boyle, A., Alton, A., Mitchell, T., Murphy, R. and Dalby, D. (2014) *Standardisation Methods, Mark Schemes, and Their Impact on Marking Reliability*. Coventry, Alpha*Plus* Consultancy Ltd for Ofqual. Available at: www.ofqual.gov.uk/documents/standardisation-methods-mark-schemes-and-their-impact-on-marking-reliability (accessed 12th February 2014).

Knoch, U. (2011) *Investigating the Effectiveness of Individualized Feedback to Rating Behavior – a Longitudinal Study, Language Testing*, 28(2), pp. 179 to 200. [no place], [no publisher]. Cited by: Boyle, A., Alton, A., Mitchell, T., Murphy, R. and Dalby, D. (2014) *Standardisation Methods, Mark Schemes, and Their Impact on Marking Reliability*. Coventry, Alpha*Plus* Consultancy Ltd for Ofqual. Available at: www.ofqual.gov.uk/documents/standardisation-methods-mark-schemes-and-their-impact-on-marking-reliability (accessed 12th February 2014).

Lamprianou, I. (2004) *Marking Quality Assurance Procedures, Identifying Good Practice Internationally.* University of Manchester. Manchester.

Meadows, M. and Billington, L. (2005) *A Review of the Literature on Marking Reliability, Report to NAA.* [no place], National Assessment Agency. Available at: https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf (accessed 6th February 2014).

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 6th February 2014).

Murphy, R. J. (1978) *Reliability of Marking in Eight GCE Examinations*, *British Journal of Educational Psychology,* 48, 2, pp. 196 to 200. Cited by: Meadows, M. and Billington, L. (2005) *A Review of the Literature on Marking Reliability*. *Reliability, Report to NAA.* [no place], National Assessment Agency. Available at: https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf (accessed 6th February 2014).

Murphy, R. J. (1982) *A Further Report of Investigations into the Reliability of Marking of GCE Examinations*, *British Journal of Educational Psychology,* 52, pp. 58 to 63. [no place], [no publisher]. Cited by: Bramley, T. and Dhawan, V. (2011) *Estimates of Reliability of Qualifications*. In *Ofqual's Reliability Compendium* (chapter 7). Coventry,

Ofqual. Available at: www.ofqual.gov.uk/standards/research/reliability/compendium (accessed 6th February 2014).

Myford, C. M. and Wolfe, E. W. (2009) *Monitoring Rater Performance Over Time: A Framework for Detecting Differential Accuracy and Differential Scale Category Use*, *Journal of Educational Measurement*, 46, 4, pp. 371 to 389. [no place], [no publisher]. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2009.00088.x/abstract (accessed 6th February 2014).

Ofqual (2011) *Inquiry into Examination Errors Summer 2011: Final Report*. Coventry, Ofqual. Available at: www.ofqual.gov.uk/files/2011-12-20-inquiry-into-examination-errors-summer-2011-final-report.pdf (accessed 12th February 2014).

Ofqual (2013a) *Review of Quality of Marking in Exams in A levels, GCSEs and Other Academic Qualifications - Findings from Survey of Examiners, May 2013*. Coventry, Ofqual. Available at: www.ofqual.gov.uk/documents/review-of-quality-of-marking-in-exams-in-a-levels-gcses-and-other-academic-qualifications-findings-from-survey-of-examiners-may-2013 (accessed 12th February 2014).

Ofqual (2013b) *Ofqual's Reliability Compendium*. Coventry, Ofqual. Available at: www.ofqual.gov.uk/standards/research/reliability/compendium (accessed 12th February 2014).

Ofqual (2013c) *Review of Quality of Marking in Exams in A levels, GCSEs and Other Academic Qualifications – Interim*. Coventry, Ofqual. Available at: www.ofqual.gov.uk/files/2013-06-07-review-of-quality-of-marking-in-exams-in-a-levels-gcses-and-other-academic-qualifications-interim-report.pdf (accessed 12th February 2014).

Ofqual (2013d) *Enquiries about Results for GCSE and A level: Summer 2013 Exam Series*. Coventry, Ofqual. Available at: www.ofqual.gov.uk/ofdoc_categories/statistics/enquiries-about-results (accessed 12th February 2014).

Opposs, D. and He, Q. (2012). 'Estimating the reliability of composite scores'. In *Ofqual's Reliability Compendium*, pp. 523-555. Office of Qualifications and Examinations Regulation: Coventry, UK .Available at www.ofqual.gov.uk/standards/research/reliability/compendium/ (accessed 10th February 2014).

Oxygen (2014) *Ofqual Quality of Marking Qualitative Research Study Final Report January 2014*. Coventry, Ofqual. Available at: www.ofqual.gov.uk/documents/ofqual-quality-of-marking-qualitative-research-study (accessed 12th February 2014).

Pinot de Moira, A. (2003) *Examiner Background and the Effect on Marking Reliability.* Manchester, AQA.

Pinot de Moira, A. (2011) *Why Item Mark? The Advantages and Disadvantages of E-marking*. Manchester, AQA, Centre for Education Research and Policy. Available at: https://cerp.aqa.org.uk/sites/default/files/pdf_upload/CERP-RP-APM-23032012.pdf (accessed 7th February 2014).

Pinot de Moira, A. (2013) *Features of a Levels-based Mark Scheme and Their Effect on Marking Reliability*. Manchester, AQA, Centre for Education Research and Policy. Available at: https://cerp.aqa.org.uk/research-library/features-levels-based-mark-scheme-effect-marking-reliability/how-to-cite (accessed 7th February 2014).

Raikes, N., Fidler, J. and Gill, T. (2010) *Must Examiners Meet in Order to Standardise Their Marking? An Experiment with New and Experienced Examiners of GCE AS Psychology,* Research Matters, 10, pp. 21 to 27. Cambridge, Cambridge Assessment. Available at: www.cambridgeassessment.org.uk/Images/109782-must-examiners-meet-in-order-to-standardise-their-marking-an-experiment-with-new-and-experienced-examiners-of-gce-as-psychology.pdf (accessed 13th February 2014).

Suto, I., Nádas, R. and Bell, J. (2011) *Who Should Mark What? A Study of Factors Affecting Marking Accuracy in a Biology Examination, Research Papers in Education*, 26, 1, pp. 21 to 52. [no place], [no publisher]. Available at: www.tandfonline.com/doi/abs/10.1080/02671520902721837 (accessed 10th February 2014).

Spear, M. (1996) *The Influence of Halo Effects upon Teachers' Assessments of Written Work*, *Research in Education*, p. 85. [no place], [no publisher].

Sweiry, E. (2012) *Conceptualising and Minimising Marking Demand in Selected and Constructed Response Test Questions*. [no place], [no publisher]. Available at: www.aea-europe.net/index.php/berlin-conference-papers (accessed 7th February 2014).

Tisi, J., Whitehouse, G., Maughan, S. and Burdett, N. (2013) *A Review of Literature on Marking Reliability Research* (report for Ofqual). Slough, National Foundation for Educational Research. Available at: www.nfer.ac.uk/publications/MARK01/MARK01_home.cfm?publicationID=948&title=A%20%20review%20of%20literature%20on%20marking%20reliability%20research (accessed 13th February 2014).

Webb, N., Shavelson, R. and Haertel, E. (2007) *Reliability Coefficients and Generalizability Theory*, *Handbook of Statistics* 26, pp. 81 to 124. [no place], [no publisher].

Wheadon, C. and Pinot De Moira, A. (2013) *Gains in Marking Reliability from Item-level Marking: Is the Sum of the Parts Better than the Whole? Educational Research and Evaluation: An International Journal on Theory and Practice*, *19(8)*, pp. 665 to 679. [no place], [no publisher].Wolfe, E. W., Matthews, S. and Vickers, D. (2010) *The Effectiveness and Efficiency of Distributed Online, Regional Online, and Regional Face-to-face Training for Writing Assessment Raters, The Journal of Technology, Learning, and Assessment*, 10(1). [no place], [no publisher]. Available at: http://files.eric.ed.gov/fulltext/EJ895959.pdf (accessed 7th February 2014).

Appendix A – Reliability of A levels and GCSEs

As part of our reliability programme, we calculated estimates of reliability for GCSEs and A levels taken in England and Wales for a number of subjects (Ofqual, 2013b).

Figure 11, based on work by Bramley and Dhawan (2012), lists some of the reliability indices at both unit level and composite qualification level for AS chemistry, AS business studies and GCSE psychology. As can be seen, AS chemistry and GCSE psychology have a composite reliability of over 0.88. However, AS business studies has a composite reliability closer to 0.8.

Regarding the acceptable level of reliability for an assessment, the general view is, when test scores are used to make important decisions about individuals in educational testing, a reliability lower than 0.80 would be considered as insufficient, between 0.80 and 0.90 sufficient, and above 0.90 good (see Frisbie, 1988; Webb *et al.*, 2007; Evers *et al.*, 2010).

Figure 11: Summary of unit and composite reliabilities for four AS and GCSE qualifications (based on Bramley and Dhawan, 2011)

Assessment	Unit	Reliability[38]
AS chemistry 3882	Unit 2811	0.813
	Unit 2812	0.827
	Unit 2813	*0.823
	Composite	0.924
AS business studies (1)	Unit 1 (winter 09)	0.641
	Unit 2 (summer 09)	0.733
	Composite	0.798
AS business studies (2)	Unit 1 (winter 09)	0.653
	Unit 2 (summer 09)	0.750
	Composite	0.819
GCSE psychology (foundation tier)	Unit 01	0.837
	Unit 02	0.839
	Unit 05 (coursework)	*0.500
	Composite	0.885
GCSE psychology (higher tier)	Unit 03	0.857
	Unit 04	0.841
	Unit 05 (coursework)	*0.600
	Composite	0.920

* Entirely or partly estimated.

[38] The overall qualification reliability was estimated using formulae for the reliability of composite scores (Feldt and Brennan, 1989; Opposs and He, 2012).

Composite qualification level reliability is calculated by aggregating multiple qualification units together. The reliability of individual units is, therefore, slightly lower. Reliability at unit level can be calculated using a measure known as Cronbach's alpha.[39] Figure 12 shows the value of Cronbach's alpha for a large number of GCSE and A level units from two large exam boards from 2009 to 2011. Figure 13 shows this information for units administered in 2012 by four of the large exam boards.

Figure 12: Distribution of Cronbach's alpha for GCSE and A level units from two exam boards, 2009 to 2011 (based on Dhawan and Bramley, 2013; Hayes and Pritchard, 2013)

Board	Type	No. of components /units	Mean Cronbach's alpha
Board A	All	287	0.813
	A level units	97	0.821
	GCSE units	190	0.809
Board B	GCSE units	142	0.83
	A level units	209	0.77

[39] Cronbach's alpha is a measure of the internal consistency of an assessment and refers to the degree to which groups of items in a test produce consistent or similar scores for individual test-takers (or consistency in test scores from different sets of items). Cronbach's alpha is affected by the quality of items in the test, the number of items, the maximum available marks for the paper and the consistency of marking of the assessment (Hutchison and Benton, 2012; He, Hayes and Wiliam, 2013).

Figure 13: Distribution of Cronbach's alpha for GCSE and A level units administered in 2012 by four large exam boards

Type	Subject	No. of components /units	Mean Cronbach's alpha
GCSE	Biology	6	0.780
	Geography	9	0.797
	French	15	0.837
	Religious studies	24	0.884
	Design & technology	12	0.856
	Physical education	8	0.830
	History	4	0.754
A level	Biology	17	0.814
	Geography	7	0.737
	French	6	0.881
	Religious studies	12	0.776
	Design & technology	6	0.810
	Physical education	5	0.828
	Maths	10	0.911

Appendix B – Review of question papers and mark schemes in 12 subjects from summer 2011 series: a summary of issues related to mark schemes

Background

In the autumn of 2011, we reviewed A level and GCSE external assessments (question papers and associated mark schemes) across a range of subjects. This was driven, in part, by the increasing number of incidents of question paper error that had been reported by exam boards during the summer series.[40]

This review was the first element of a programme of work on the quality of assessment materials. It aimed to build a picture of overall quality of assessment materials as well as identify trends in this over a number of years. This particular review was conceived and designed as an indicative and initial, but informed, review on specific aspects of quality of exams taken in summer 2011. The findings were not definitive regulatory findings. They did, however, give intelligence on the quality of assessment design.

The review focussed on eight areas affecting the quality and standard of question papers and their associated mark schemes. Of these areas, one only related to mark schemes, specifically the "ability of mark schemes to produce fair and consistent outcomes". As an indicative investigation, only one subject expert reviewed assessment materials in each subject. The focus of this work was on external assessment, and only external exam question papers and mark schemes were considered.

The qualifications and subjects reviewed were selected for having reasonable entry and not having been monitored recently. It was decided for some subjects to focus on GCSEs, for some to focus on A levels, and for some to focus on a combination of the two. For A levels, both AS and A2 units were reviewed, and for tiered GCSEs, both foundation and higher tier units. Similarly, it was decided in some cases to focus on one exam board, but in other cases to focus on two of the three England-based exam boards. Taking into consideration the fact that some qualifications are not offered by all the English exam boards, an equitable spread of exam boards' papers across 12 qualifications and subjects was achieved. This was as follows:

A level and GCSE

ICT
Design & technology

[40] www.ofqual.gov.uk/standards/inquiries/exam-paper-errors-2011

Psychology
Leisure and tourism/Leisure studies (leisure studies AS units only and leisure and tourism GCSE)
Media studies

GCSE only

Health and social care
History

A level only

Business studies
Chemistry
Economics
General studies
Sociology

In summary, there were no common significant standards or quality issues with the assessments. All assessments were deemed to be fit for purpose, but some issues were identified. Over half of these related to mark schemes: these are summarised below. A common finding was the need for quality checking of all parts of mark schemes to make sure they are in line with question papers.

Design & technology

- In one GCSE and one AS mark scheme, there were some questions where there was potential for a lack of discrimination between students' responses. The mark schemes awarded marks based on the number of points or issues raised rather than crediting the depth of understanding shown by a student.

History

- In two GCSE mark schemes, there were two questions where the mark scheme required a greater level of understanding and skill than the questions asked for. For the higher marks, these mark schemes required students to show an understanding of or analyse the key features of the questions, whereas the questions themselves only asked for a description.

Health and social care

- In one GCSE mark scheme, there was one question where the mark scheme required a greater level of knowledge and understanding than the question asked for. This mark scheme required students to give an explanation in

relation to the subject of the question, whereas the question itself only asked for a description.

ICT

- In one GCSE mark scheme, there were two questions where the mark scheme did not give enough detail to indicate how marks should be awarded. In one instance, the mark scheme did not specify the number of points students should make to attain different mark bands. In the other instance, the mark scheme did not sufficiently clarify the meaning of key words or phrases used in the mark bands.

General studies

- In one AS mark scheme, there was one question where the mark scheme penalised students by instructing examiners to deduct a mark for a wrong answer. This contradicted the general guidance on marking to apply mark schemes positively.

Psychology

In one A2 mark scheme:

- There was one question where the mark scheme required a greater level of knowledge and understanding than the question asked for. For higher marks, the mark scheme required students to give an explanation of issues, whereas the question only asked them for a description of the process they had carried out.

- There was one question where the mark scheme indicated that there were a greater number of marks available than allocated on the question paper.

In two GCSE mark schemes:

- The general guidance on marking was missing.

- There were some questions where the mark schemes did not give enough detail to indicate how marks should be awarded. For example, whether it was one mark per point made or one mark per point with development.

Media studies

In one AS and one A2 mark scheme:

- There were two whole sections, one in each paper, for which the mark bands were too wide, with the potential for a lack of discrimination between students.

- For one section of the AS paper and the whole A2 paper, the mark schemes did not contain enough indicative content to enable consistent marking.

Leisure and tourism/Leisure studies

- In one GCSE and one A2 mark scheme, the mark scheme requirements were sometimes below and/or exceeded those of questions. For example, one exemplar response was awarded full marks, but did not fulfil the question requirements and, conversely, a top mark band required a greater level of skill than the question asked for.

In two GCSE and one A2 mark schemes, some levels of response mark schemes did not contain enough indicative content to enable consistent marking.

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 2014

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