Qualifications and
Curriculum Authority

Examination standards

Report of the independent committee to QCA

Barry McGaw, Caroline Gipps, Robert Godber

December 2004

Contents

Executive summary

The Independent Committee on Examination Standards was established by the Board of the Qualifications and Curriculum Authority (QCA) on the basis of a recommendation in the final report of Tomlinson's inquiry into A level standards in December 2002.

The Committee was appointed in late 2003 and comprises:

- Dr Barry McGaw, Director for Education, Organisation for Economic Co-operation and Development (OECD) – Chair;
- Professor Caroline Gipps, Deputy Vice-Chancellor, Kingston University;
- Mr Robert Godber, former Headteacher, Wath upon Dearne Comprehensive School, Rotherham.

The remit was to focus on 'examination standards' and QCA's regulatory role with a specific focus on GCE A levels.

The Committee's conclusions are:

- No examination system at the school or other level is so tightly or carefully managed.
- Strategies for maintaining comparable examination standards across awarding bodies are adequate to the task.
- Strategies for maintaining comparable examination standards across time do as well as possible, but there are unrealistic expectations still in play.
- Strategies for determining whether comparable performance standards are being maintained are difficult to strengthen in the environment of public examinations.
- No examination system has found an adequate way to determine whether standards are constant across subjects.
- The English public and its media expect too much of the public examination system at school level.
- The awarding bodies have broadly consistent and well-regulated systems for setting question papers, managing marking and awarding grades.
- QCA has robust systems in place to monitor and regulate the work of the awarding bodies.
- Implementation of electronic distribution of student examination answers at the question level will improve control of marking and monitoring of markers.

- Limited progress has been made in addressing the research questions proposed by Baker et al. (2002) and we question whether sufficient resources have been allocated to this work.

- Consideration should be given to commissioning awarding bodies and other agencies to carry out some research tasks within a framework of a coherent and phased research programme.

- QCA has undertaken an exhaustive analysis of risks, and the establishment of the NAA to handle delivery issues is a welcome development.

- QCA has substantially improved its performance in communicating with professionals and the public about its procedures. It should continue to be more proactive in public discussion, based on good quality data and rigorous evidence of the efficacy of the procedures that it and the awarding bodies use. To the extent possible, QCA's voice should be distinguished from that of the government of the day to preserve and strengthen its role in the discussion of examinations policy and practice.

Much of the public discussion of examination results in England is based on the assumption that results are standard-referenced with a degree of precision that cannot be delivered. Over the longer term, it makes little sense in many subjects to ask whether examination standards have been maintained since the subjects themselves have changed so much. It would help if different expectations were set, not asking if performance standards are rising or falling over the long term but asking only if the examinations are making reasonable and appropriate demands of students and if the results work for the key purposes for which they are intended.

The committee's overall conclusion is that no examination system at the school level is better managed. Most countries have self-regulated government departments or agencies conducting examinations. In England there are three large and highly professional, independent awarding bodies operating within a tightly regulated environment with many checks and balances.

This system is not without its problems, of course, some of them unintended or inadequately anticipated consequences of related decisions. The most notable example is the essentially simultaneous introduction of a modular course structure and AS courses, both of which increased enrolments and examinations in a system already under stress.

Expectations of examination systems

Discussions of whether standards are being maintained are made more complex by the multiple purposes that examination systems serve.

Standards in education

Debates about whether standards in education are being maintained often founder because adequate information with which to answer the central question is not available. They also founder because the term 'standards' is used in different ways.

'Standards' can refer to the demands the education system places on students. These standards are expressed in the curriculum, in terms of the breadth and depth of learning required. These *curriculum standards* define what students are expected to know and be able to do. Standards are also expressed in the tasks that students are required to perform in school-based assessment and external examinations. These *examination standards* provide students with a particular opportunity to demonstrate what they know and are able to do. If the school-based assessments and examinations are valid, the examination standards will be consistent with the curriculum standards.

'Standards' can also refer to the levels of learning students actually achieve, that is what they actually know and are able to do. These can be called *performance standards*.

Examination standards and performance standards interact, of course, and this can result in irresolvable differences in interpretation of the same evidence. As Stobart (2000) points out

> This ambiguity [in the meaning of 'standards'] leads to the August ritual of any improvements in the GCSE/GCE pass rate being welcomed by some as an improvement in [performance] standards and denounced by others as further evidence of falling [examination] standards (p. 3).

In England, debates about whether performance standards are rising or examination standards are falling plague discussions of the results in the key stage assessments and in the public examinations at GCSE and GCE. As some vocational qualifications are brought within the GCE framework, debate about the meaning of awards is beginning to sharpen there as well.

University assessments of students seem largely to escape this kind of public debate, despite the very limited four-point scale on which final honours results are expressed and in the absence of any commonality of assessments across institutions, even within the same fields of study.[1]

[1] Commonality of standards is sought through national subject benchmark statements and the use of external examiners who report nationally on the maintenance of academic standards. Brazil is the only country to have introduced common examinations across

It is interesting that the debate is much more intense about the GCE results on which universities depend in large part for their selection of students than it is about the university results on which employers depend in large part for their selection of staff from among university graduates. There are obvious enough reasons for this – among them, the relatively small number of universities compared with the large and diffuse range of employers; the sense (not necessarily well-founded) that the transition into university is more life-determining than the initial transition from university into work, and the more public nature of the selection at entry to university than at entry to the workforce.

In this report on examination standards, we focus on GCE A level examinations.

While the debate is generally about whether examination standards and/or student performance standards are being maintained at the same level over time, there are other important questions as well.

One is whether both examination and performance standards are comparable across Examination Boards (the three in England and those in Wales and Northern Ireland and, perhaps, that in Scotland as well since students move between those jurisdictions between school and university).

A further, more difficult question is whether examination standards and performance standards are comparable across subjects.

Later in this report, we discuss the strategies used by the Boards and their regulators to address these questions. The focus of this report is on the awarding bodies based in England and the Qualifications and Curriculum Authority (QCA).

Another important question, which we do not address since it is not related to standards in examinations, is whether the demands in the curriculum are set at an appropriate level. Judging whether they are involves consideration of the characteristics of the students who enrol, the purposes for which they do so and the uses to which they and others expect them to put the knowledge and skills they acquire.

Monitoring population performance standards over time

If the only task were to determine whether performance standards in the student population were rising, remaining constant or falling, then it would be relatively straightforward to do it satisfactorily, at least in some circumstances.

For an age-group in which everyone is still at school and a subject area that all students study, determining what is happening to performance standards would require testing of only an appropriate sample of students. Sufficient assessment materials could be kept confidential after one testing for re-use in one or more subsequent years to link the tests on different occasions. This would enable

---

universities for particular fields of study but that system did not survive the most recent change of government.

results from successive years to be expressed on the same scale, and so to hold the examination standard constant.

Holding the examination standard constant in this fashion does not require the difficulty of the examination (or test) to be precisely constant from one year to the next, since that can never be guaranteed. It is the scale on which students' results are expressed that is held constant. That is achieved by statistically adjusting for unintended fluctuations in the difficulties of examinations (or tests) over time. Using the same scale on which to express students' results permits direct comparisons of results in different years to determine whether performance standards are changing.

When only a sample of students is used to determine what is happening to the population as a whole, the assessments are essentially 'low-stakes' for individual students, teachers and schools. There is, therefore, little incentive for teaching to the test in a way that would invalidate the performance measures. A greater threat to the validity of the assessments could be that the tests are not taken sufficiently seriously, with the consequence that the sample of students underperforms.

There are national and international examples of surveys of this kind, that hold the examination standard constant in the manner described and assess samples of students on a common scale to determine whether performance standards change over time. They include the former English Assessment of Performance Unit (APU) surveys, the US National Assessment of Educational Progress (NAEP) and the OECD Programme for International Student Assessment (PISA). These surveys are based on frameworks that define what students are expected to know and be able to do (examination standards) and they assess what students actually know and are able to do (performance standards). These 'examination standards' are defined in more general terms than those for which there are detailed subject syllabuses with relatively precise expectations of what students are to learn, as in GCE A level courses.

Satisfactory links between assessments of successive samples of students can be sustained over a number of years only if the nature of the knowledge and skills to be assessed remain essentially the same. If they change in ways that alter the test framework and require changes in the tests themselves, it can become impossible to maintain a common scale. In that case, the examination standard would have altered and students' performance standards could no longer be compared over time.

Measuring individual student performance standards

Often interest is not only in whether the performance standards of the population are changing over time but also in the current performance levels of individual students and, perhaps, their schools. In that case, it is necessary to test all students and not just a sample.

If comparisons to be made are only among students taking the same assessment, it would not matter if the examination standards (and thus the tests or examinations) changed over time. Indeed, it could be quite proper for them to

be changed deliberately to reflect changes in the knowledge or skills base to be assessed or changes in the kinds of students studying the curriculum and taking the examinations.

This could clearly be the case with GCE A level curricula and examinations. There have been marked changes in many subject areas, perhaps most in science and technology. If the purpose is to judge how well students perform in physics for example, either in comparison with each other or in comparison with the current examination standard, it would not matter if the current examination standards in physics were different from those of some years or decades earlier.

Comparing students' performances with those of other students is traditionally called 'norm-referencing' to indicate that students are compared with the average performance or norm. If the whole age population is involved, it could be called 'population-referencing'. If only a subset of students is involved, as is always the case with GCE A level subjects, then it would more appropriately be called *cohort-referencing*.

Comparing students' performances with the examination standard involves what could be called *standards referencing*, though it has typically been called 'criterion-referencing'.

While 'cohort-referencing' and 'standards-referencing' differ markedly in purpose, they do not require fundamentally different kinds of examination. The difference occurs in the way in which the performance results are interpreted and used.

Some examinations systems opt for this limited form of comparison, time bound within the year of the examination. In its simplest form, a fixed percentage of candidates receives each grade. Grades are interpreted to have the same meaning over a relatively small number of years on the assumption that the cohort of students does not alter quickly. Over a longer period, if the cohort changes because of an increase in the proportion of the population participating or because of shifts in the subject enrolment preferences of students, grades awarded in this way cease to be comparable.

Seeking to do both at once

Some examination systems seek to compare students with one another and against the current examination standard while also seeking to compare performance standards over time.

GCE A level examinations are one example. Students are awarded marks and grades (A to E and U) with which to make comparisons among the students in a particular year. In addition, attempts are made to ensure that grades, and to some extent the marks behind them, are comparable over time.

If some parts of an examination can be kept confidential and re-used in later examinations, it is relatively easy to establish links between examination scales from year-to-year, and so to monitor any changes in performance standards over time. The *Graduate Australian Medical Schools Admissions Test* provides

an example. Results from tests taken by candidates in different years can be compared directly since the examination standard has been held constant through the use of a constant scale on which results are expressed. Performance standards could alter quite markedly over time if the type of candidates changes or the quality of the education that candidates have received changes. Such real differences in candidates from year to year would be reflected in differences in scores on the constant scale used for examination results.

With public examinations like GCE A levels, there is a need for an open disclosure of the contents of each examination. Material cannot be kept confidential after use to be used again in order to locate subsequent examination results on a common scale that holds examination standards constant and permits performance standards from different years to be compared. Examiners may work diligently to maintain a consistent examination standard by trying to set questions of comparable difficulty in successive years but there is as much art as there is science to this. Without the capacity to repeat some questions and to build links between the scales on which marks are awarded across years to keep them in line, achieving comparable difficulty across years is essentially a matter of experience and professional judgment.

One system that attempts to do this is the end-of-secondary-school Higher School Certificate examinations in New South Wales in Australia (McGaw, 1997). In the first stage, experienced markers inspected examination papers and marked students' work from a prior year and developed descriptions of student performance at five bands (equivalent to A to E) with a further, un-described lower category for inadequate work. From the following year these band descriptors, together with marked student work at grade boundaries in past examinations, have been used to establish where to set grade boundary marks on each question and thus on the paper as a whole (Bennett, 2001). By this means, careful attention is given to applying the same examination standards in successive years.

Examination standards can change because of changes in the curriculum. They can also change because of changes in examinations themselves. In many jurisdictions, the examination process has become both more open and more explicit. Examination questions make much clearer to students than in the past exactly what is required of them. Examiners report to teachers on the overall performance standards of students in ways that help teachers learn how best to prepare their students for the examinations. That is clearly the case with GCE A level examinations. To the extent that these improvements in the examination process improve student performances, it is a consequence neither of an improvement in performance standards nor of a decline in examination standards but rather of improvement in the validity of the examinations. Examinations that are difficult because students have trouble working out what the examiner wants do not measure performance against the examination standard.

Alongside all the attempts to relate performance standards to examination standards, there is the task of making comparisons among students. This cohort-referencing is important when selections among students have to be

made, most obviously by universities. Particular problems have arisen in England in recent years as the A-grade in the GCE A level courses has ceased to provide sufficient differentiation among students for some particularly competitive selections to be made.

The question of whether to use the actual marks on the examination scale or bands of marks to create categories such as grades is important for all comparisons, both those among students and those over time. Categories, like grades, simplify the performance message but they create two problems. One is that they distort comparisons among students to some extent. A student close to the bottom of the range of marks for an A-grade is much more like a student close to the top of the range for a B-grade than to a student at the top of the range for an A-grade. The grade awards, however, declare the two with As to be the same and both similarly different from the student with the B. The second problem is related. There is always some imprecision in measurement, including in examinations. This matters most at grade boundaries, where misclassifications are more significant than imprecision in the underlying marks themselves. Even if the marks are made public, imprecision in them matters less because the differences are smaller and so less significant.

The task of monitoring performance standards over time, with a constant examination standard, is feasible. The tasks of measuring performance standards against an examination standard at a particular time, and making comparisons among students, are feasible. To do all of this at the same time asks a great deal of an examination system. The key question is whether it is an attempt to do too much that will, in the end, mean that none of the tasks is done as well as it might be.

The examination process

In this section, we describe the processes used in setting, marking and grading assessments of students' work in England, Wales and Northern Ireland. Although this information is publicly available on the website of the Qualifications and Curriculum Authority (QCA), we suspect that the details are not widely known or well understood. We believe that any serious public discussion of the quality and utility of the examination system must be based on a full and fair understanding of its processes.

We focus on the GCE A level examinations since they are the ones about which public discussion and debate are usually the most strenuous.

Apart from Scotland which has a different education system, there are five bodies in the rest of the United Kingdom that provide curricula, conduct examinations and award general qualifications at the secondary school level: the Assessment and Qualifications Alliance (AQA), Edexcel and the Oxford Cambridge and RSA Examinations (OCR) based in England, the Welsh Joint Education Committee (WJEC) and the Council for Curriculum, Examinations and Assessment (CCEA) in Northern Ireland[2]. Schools and colleges in all three countries have access to the qualifications offered by all five awarding bodies.

Nature of assessments

Components of GCE A level subjects are assessed either by an external examination taken at a specific time or on the basis of work completed over a longer period of time. The latter is typically assessed in the school or college and is usually referred to as 'coursework'. The permissible balance of internal and external assessment is prescribed for each subject in the specification.[3]

Examinations

For examinations, a principal examiner drafts a paper and mark scheme which is then reviewed by another for coverage of the specification for the subject, comparability with previous papers, clarity and so on. The paper is revised and then reviewed by a Question Paper Evaluation Committee under the Chair of Examiners for the subject. That committee may suggest further amendments. After any further revision, the paper goes to an assessor or scrutineer who checks that it is fair to candidates and can, for example, be completed in the time allowed. The Chair of Examiners signs off on the final version.

Coursework

Coursework tasks may be set by an awarding body, by teachers and then approved by an awarding body or by teachers according to guidance in the specification for the subject. Coursework tasks, like examinations, are designed

---

2 Earlier bodies were generally referred to as 'examination boards'. They are now referred to as 'Awarding Bodies' and that nomenclature is used in this report.

3 A 'specification' provides the course syllabus and details of assessments, including marking criteria that are to be used.

to assess students' performances against assessment criteria set out in the specification for the subject.

Marking students' work

Marking completed examination papers

Before marking of students' scripts commences, the principal examiner convenes a standardisation meeting with markers to ensure they all interpret the mark scheme in the same way and will deal similarly with any problems that have by then been identified with the paper. At this meeting, the markers also mark a number of common scripts and review their marks to confirm that they are working consistently.

At regular intervals during the marking process, samples of examiners' marking are checked by more senior markers to ensure that they are maintaining their consistency. If there is evidence that a marker has not been marking in line with the required standard, then the marker is required to adjust his or her marking to bring it into line. In extreme cases, the marker will be stopped from marking and the scripts will be re-marked.

Under the standard procedures, batches of examination papers are sent from schools and colleges to examiners designated by the awarding body. Direct despatch from schools and colleges speeds the process but it has some disadvantages. First, it means that all papers from a school or college are marked by the same marker. Secondly, the marker may well be able to identify the school or college from which the papers have come. There is no direct identification of the school or college on the papers or the parcel containing them but the postmark on the envelope sometimes reveals this information. To mitigate any potential for bias all markers must declare an interest in schools or colleges with which they have a close association through, for example, having taught there recently or having a relative there as a student. They are not allocated students' scripts from such schools or colleges.

Some examination scripts are now being scanned and distributed electronically to markers who then read the images of the students' hand-written responses on their computer screens. This speeds up the distribution process but it also means that markers can be sent answers to single questions rather than whole scripts. This ensures that individual students' total scripts are not marked entirely by a single marker; that markers' attention is focused on fewer questions and so more likely to be consistent; and that clerical staff rather than examination markers can mark any questions for which answers are unambiguously correct or incorrect and so reduce the demand for qualified markers. While no marker then looks at a student's complete examination script to form an overall view, that actually has the advantage of reducing the risk of a 'halo effect' where a good answer to one question leads an examiner to take a more lenient view of a poor answer to another, or vice versa.

There is some evidence that marking of single questions on the screen in this manner is slightly harsher than marking questions within whole scripts in their original paper form. If both types of marking are used for a single examination

there is a danger of putting students whose scripts are 'e-marked' at a disadvantage. Either one or the other type of marking should be used for all scripts in an examination, unless a way of adjusting for the difference is established.

Marking coursework

Teachers within a school or college assess their students' work against the criteria. Their marks are submitted to the awarding body and a sample of marked students' work is sent to a moderator, or the moderator visits the school or college depending on the nature of the coursework. The moderator checks that marks have been awarded in line with the agreed national standard.

Moderators, at this stage, will have already under-gone standardisation to ensure they have a common understanding of the mark scheme. If the original marks from the school or college are consistent with those of the moderator within an agreed tolerance, the original marks are accepted. If, on the other hand, the original marking is out of line, the moderator then marks the entire sample and the marks are analysed to determine whether all of the marks from the school or college need to be adjusted. This strategy is designed to ensure that the marks awarded by a school or college are in line with the agreed national standard[4].

Moderators' work is checked at regular intervals during the moderation process by senior moderators to ensure that their judgments are consistent and in line with the agreed standard.

Marking markers

In addition to the monitoring of individual markers and moderators during the marking process to provide advice on their judgements or to adjust their marks, all of the awarding bodies make an overall assessment of each one of them. They are given performance grades, for example from A to E, and are given written reports including their grade.

Markers and moderators given an E-grade are not used again. Some who might have received an E-grade are stopped during the processes and so do not continue even to the end of that marking session. They also are not used again. Those receiving a D-grade may be used again but with a reduced work-load and only if they undertake re-training. In recent years, fewer than 3 per cent of markers and moderators have received and E-grade and around the same percentage have received a D-grade.

Awarding results

Determining grades

When all the examination scripts in a subject have been marked, senior examiners and a staff member from the awarding body involved with the

[4] Moderation is a complex process and there is some variation among the awarding bodies in the precise procedures that they follow, within the framework of the Code of Practice.

examination are convened by the chair of examiners for the subject to recommend boundaries for the grades (A and E for GCE A level subjects). They review examination scripts with marks in the region of the ones proposed by the principal examiner for the paper for the A/B and E/U grade boundaries. They also review completed scripts from previous years that were close to these two grade boundaries. The principal purpose is to maintain the examination standard for these two important grade boundaries from year-to-year. For a GCE subject, these meetings typically take two days.

In practice, the process typically establishes a small range of marks within which the meeting is unable initially to decide precisely where to locate a boundary. The meeting seeks to base its judgement on evidence in the scripts from the current series and in comparison with archive scripts from previous years. This establishes a 'zone of uncertainty'. The boundary then needs to be fixed somewhere within this range of marks.

The meeting will also consider the distribution of results that would be produced, comparing the percentages of students who would achieve an A-grade and a pass grade (E or better) at particular boundary marks with the corresponding percentages from previous years. The meeting is provided with any evidence of changes in the cohort of students taking the examination over the years in question that might explain and justify any marked shift in the percentages of students achieving particular grades. This information can include achievement at GCSE for the students involved or results in other A level examinations. Using all of the evidence available, the awarding committee will choose the single mark which they recommend to the Chief Executive of the awarding body as the lowest mark which is worthy of the grade (A and E in the case of GCE A levels).

While much media and public attention on the release of results is given to any changes in the percentages of students receiving particular grades, there is usually no way to maintain an exact match even if that were wanted.

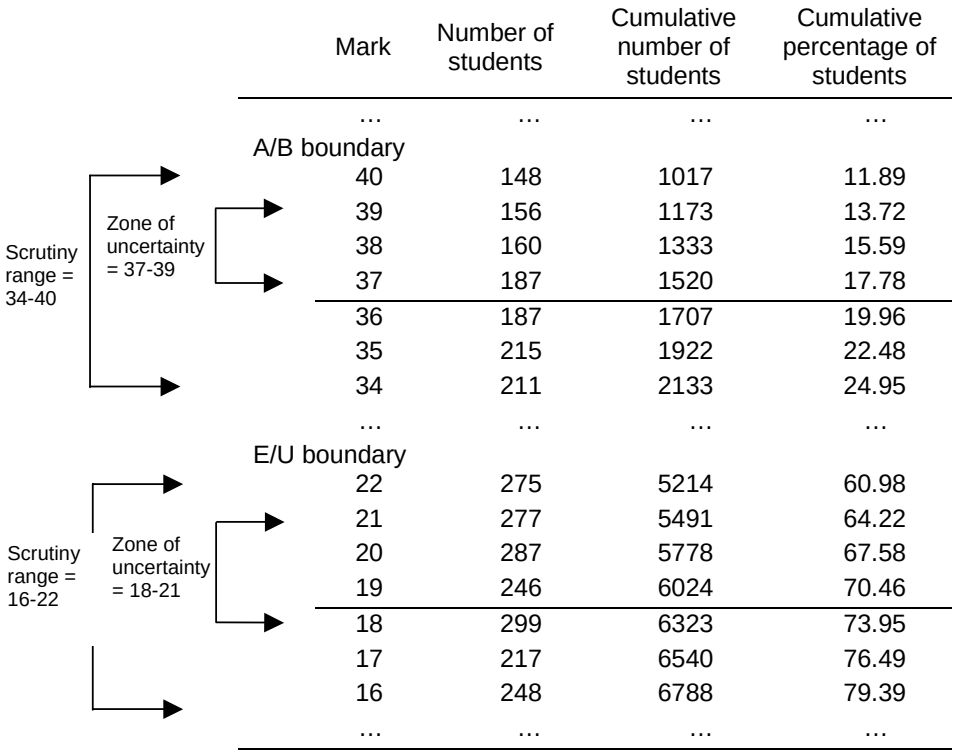
The actual results from one examination paper in one awarding body, shown in Figure 1, illustrate the point. In this case, the awarding committee In this case, the awarding committee examined papers with marks in the range 34-40 in their consideration of where the A/B boundary should be located. This 'scrutiny range' is established in advance by the Chief Examiner and staff of the awarding body who make sure that there are several marked scripts available in the meeting for each mark in the range. In considering which marks represent performance at A-grade and which at B-grade level, the participants in the meeting also review marked scripts from previous years that were either side of the A/B boundary.

After reviewing the scripts the meeting narrowed the range of marks within which the boundary could be set to 37 to 39 and this constituted their 'zone of uncertainty'.

In 2003, 18.25 per cent of students in the equivalent examination paper had been awarded an A. The meeting in 2004 positioned the boundary between 37 and 36 and thus awarded As to 17.76 per cent of students. They could not have

matched the 18.25 per cent of the previous year even if they had wanted to. The clustering of students on each permit it.

For that E/U boundary, the scrutiny range was 16-22 and the zone of uncertainty was 18-21. In 2003, 73.12 per cent of students had received grades of E or better. In 2004, the committee set the boundary between 19 and 18, thus awarding grades of E or better to 70.46 per cent of students. The meeting could have set the boundary between 18 and 17, since the mark of 18 was also in the zone of uncertainty, but the judgement was that scripts on 18 were not worthy of a grade E, compared with the previous year.



Mark	Number of students	Cumulative number of students	Cumulative percentage of students
…	…	…	…
A/B boundary			
40	148	1017	11.89
39	156	1173	13.72
38	160	1333	15.59
37	187	1520	17.78
36	187	1707	19.96
35	215	1922	22.48
34	211	2133	24.95
…	…	…	…
E/U boundary			
22	275	5214	60.98
21	277	5491	64.22
20	287	5778	67.58
19	246	6024	70.46
18	299	6323	73.95
17	217	6540	76.49
16	248	6788	79.39
…	…	…	…

Figure 1: Examples of boundary locations for one examination paper in 2004

The use of the expression ‘zone of uncertainty’ makes clear that the senior examiners in the meeting do not reach a precise judgement on student performances in relation to the examination standard and so of where the grade boundary should be set. Their final judgement in each case determines, in part, how the percentage of students receiving a particular grade relates to the percentage from the previous year. The process is thus criterion-referenced or standards-referenced in the first stage and then, in the final stage, both norm-referenced to the extent that past grade distributions guide the final choice and standards-referenced to the extent that comparisons with scripts from previous years guide the choice.

Review of grading

Once the awarding committee has decided on its recommendations on the A/B and E/U boundaries, the boundaries between the intervening three grade pairs

are interpolated to generate the full distribution of all grades for the subject. The results are then reviewed in two stages.

First, the chief executive of the awarding body or his or her representative reviews the awards, considering any issues that the awarding committee has raised and taking account of external information such as results in other subjects and results in the same subject from other awarding bodies. The boundaries between grades can be moved at this point but with the chair of examiners' agreement and not normally outside the 'zone of uncertainty' established in the meeting of examiners.

Once the grade boundaries have been finalised, a team of senior examiners meets to re-mark the scripts of any candidates who are considered to be at risk of receiving the wrong result. This may be, for example, because their final grade differs markedly from an estimate provided in advance by their school or college. This review normally involves only the examination components because the bulk of coursework is held by schools and colleges, not the awarding body.

Conversion of marks to uniform mark scale

The actual marks awarded in a particular assessment (examination or coursework) are not directly comparable across assessments within a subject or across years. They depend on the difficulty of the particular assessment and the marking scheme used. As a final step, marks are converted to a uniform mark scale.

On the uniform mark scale the lowest mark for an A is set at 80 and the lowest mark for an E is set at 40%. The actual marks from an assessment are converted to the uniform mark scale by:

- converting the minimum mark for A to 80% of the total available;
- converting the minimum mark for E to 40% of the total available;
- converting marks between the two minima by simple ratio to marks between 40 and 80 on the uniform mark scale;
- converting marks that obtained an A to the range from 80 to 100;
- converting marks below an E to the range from 0 to 39.

If the original minimum for an A is below 80 and the original minimum for an E is above 40, the method of converting original marks in the A range and the below-E range to the uniform mark scale is adjusted to compensate for this bunching in the middle of the range of original marks. A mark on the original scale above the minimum mark for an A by twice the range of marks for a B is converted to 100 on the uniform mark scale. A mark on the original scale below the minimum mark for an E by the range of marks for a B is set at 30 on the uniform mark scale.

The uniform mark scales are then essentially comparable across all the component assessments for a subject and can be used, in addition to the grades for individual components, when overall grades for the subject are being determined for each student.

Enquiries about results and appeals

Students dissatisfied with their results can, through their school or college, request re-marking of their examinations scripts or coursework. The outcome can be to leave a grade unaltered, to raise it or lower it. Students dissatisfied with the outcome may appeal against the process to the Examination Appeals Board from which a final ruling, not subject to further appeal, will be delivered.

Quality assurance and control mechanisms

The systems employed by the awarding bodies for setting assessment tasks for examinations and coursework and for marking and grading student work are elaborate and include considerable independent checking. Beyond those systems, there are formal quality assurance and controls processes managed by regulatory authorities. We turn now to an analysis of these.

The three GCE A level awarding bodies in England – AQA, Edexcel and OCR – are regulated by the Qualifications and Curriculum Authority (QCA); WJEC is regulated by the Qualifications, Curriculum and Assessment Authority for Wales (ACCAC); and CCEA is a self-regulatory body in Northern Ireland[5].

The regulatory authorities draw up criteria for the accreditation of qualifications to the national qualification framework. These cover areas such as general principles, qualification-specific issues and subject-specific content. The awarding bodies then draw up specifications (syllabuses) to meet these criteria.

In describing the quality assurance and control mechanisms of the regulatory authorities, we focus on QCA as the largest, but it works closely with the others in developing common strategies and in carrying out some of its monitoring work.

QCA has a planned programme of quality assurance and control activities which is reviewed regularly. The purpose of this monitoring work is to promote public confidence in the quality of external qualifications by ensuring awards meet the regulatory requirements for quality, rigour, fairness and consistency and that awarding bodies are delivering particular qualifications according to the accreditation criteria and code of practice. QCA also uses information gathered through monitoring activities to review its own practice, such as improving accreditation procedures.

Quality assurance mechanisms

Code of practice

One of the key elements in QCA's quality assurance programme is the *Code of Practice*, which applies to GCSE, GCSE in vocational subjects, GCE, VCE, GNVQ and AEA qualifications. The code is produced by QCA in collaboration with the regulators in Wales and Northern Ireland and both prescribes and guides the practices of the awarding bodies.

The Code is designed to promote quality, consistency, accuracy and fairness in assessment and awarding and takes into account changes in government policy in relation to the relevant qualifications. The code is intended to ensure that standards are maintained in each subject, across awarding bodies and different

[5] The formal provision is that each regulator has responsibility for the operations of the schools and colleges within its jurisdiction. In practice, they co-operate to avoid duplication of effort.

syllabuses from year to year. The intention is that all candidates should receive a fair result, regardless of the general qualification they are taking, the awarding body delivering the qualification or when they are assessed. The code sets out:

- agreed processes and practices for the assessment and quality assurance of GCSE, GCSE in vocational subjects, GCE, VCE, GNVQ and AEA qualifications;
- the responsibilities of awarding bodies, schools and colleges and regulatory authorities for the provision and administration of assessment;
- the basis upon which the regulatory authorities will systematically monitor the performance of awarding bodies in maintaining the quality and standards across the accredited qualifications they offer.

The code is reviewed regularly and changed in the light of:

- awarding bodies' experience of using the code in practice;
- research on best practice in assessment and awarding;
- evidence obtained by the regulatory authorities on the performances of awarding bodies.

QCA's observations of awarding meetings and discussions with the awarding bodies had revealed some differences in the interpretation and application of the code's specifications about how information was to be used in setting grade boundaries. In response, the wording of the code (paragraph 128) was clarified in order to achieve greater consistency of practice across awarding bodies.

The code is also altered to reflect improved capacity in the awarding bodies. For example, evidence that the bodies were able to provide the outcome of a priority enquiry about a subject grade faster than the deadline of 30 days resulted in a revision of the required response time in the code to 20 days.

Senior examiners' conferences and technical seminars

As part of its quality assurance programme, QCA and its fellow regulatory authorities hold seminars and conferences with representatives of the awarding body. These aim to improve the quality of external qualifications by investigating problems and disseminating good practice. In autumn 2003, for example, QCA hosted a technical seminar and three senior examiners' conferences.

Technical seminars typically involve an evaluation of particular aspects of the current examining system as well as consideration of forthcoming changes to the system or the technical issues surrounding a new qualification. The autumn 2003 seminar focused on the implementation of a paragraph of the Code of Practice which is at the heart of the awarding process and on the use of the uniform mark scale in the revised GCSE examinations.

Senior examiners' conferences span the main GCSE and A level subjects. The conferences give participants the opportunity to share good practice across subjects, as well as engage in detailed discussion of subject-specific issues. Feedback from participants suggests that these activities are valuable.

Self-assessment

QCA requires awarding bodies to have systematic arrangements for monitoring and reviewing all aspects of their own work and discusses with them the development of their practices for doing so. As part of this process, the awarding bodies are required to develop action plans to promote continuing improvement.

In 2003, QCA published *Producing self-assessment reports: Guidance for awarding bodies* in which it:

- synthesised the requirements of the various regulatory codes under 12 key statements or reporting areas;
- provided general guidance on the development of self-assessment reports.

The self-assessment reports, which are submitted by the awarding bodies to QCA, document both problems and progress. One awarding body's report for June 2002 recorded that marks for 91 per cent of candidates were available at the awarding meeting. The report for June 2003 recorded that marks for 95.5 per cent of candidates were available.

Quality control mechanisms

QCA's monitoring programme also includes a range of quality control activities. This involves monitoring of processes and standards. Strategies used to monitor processes include 'scrutinies', scrutiny follow-up and code monitoring. Strategies used to monitor standards include reviews of standards, cross-subject and cross-qualification comparability studies and 'probes'.

Monitoring of processes

Scrutinies

Each year QCA carries out up to 30 'scrutinies', which are detailed checks of individual GCSE or A level examinations. The selections of examinations for scrutiny are based on judgements of risk, taking account of how recently the subject has been accredited, the size of the enrolment for the examination and, sometimes, information that there may be particular problems. Selections are also made to ensure there is an appropriate distribution of scrutinies across awarding bodies and, over time, appropriate coverage of subjects.

QCA has appointed a pool of independent subject specialists with appropriate teaching and assessment experience to lead the scrutiny work. Additional subject experts are commissioned to work on the analysis of assessment materials and candidate work.

Scrutinies involve observation and reporting on awarding body meetings held at various stages of the examining process as well as detailed analysis of question papers, mark schemes and candidates' work. The purposes of a scrutiny are to:

- determine whether the required national criteria or principles and associated codes of practice have been met;

- determine whether the assessments were fair and effective in measuring candidates' achievements in relation to the stated assessment objectives;
- determine whether the procedures designed to secure consistency of practice and comparability of standards were implemented effectively;
- identify any aspect of the syllabus which appears to have constrained fair and effective examinations;
- identify good practice worthy of encouragement and dissemination;
- provide a report to the awarding body on the findings of the scrutiny, including recommendations on issues to be addressed to improve performance.

On receipt of a scrutiny report, the awarding body must produce an action plan to be submitted to QCA showing how it intends to address each of the recommendations. QCA then decides what evidence will be required to establish that each issue has been addressed satisfactorily. This may be documentary evidence or involve observation of a meeting. When all action has been completed successfully, QCA advises the awarding body that it has 'signed off' on the action plan.

One 2003 scrutiny included a recommendation that the particular awarding body ensure that, in meetings to standardise work and set grade boundaries, there are sufficient scripts to illustrate the range of performances. This was followed up by observation of the May 2004 standardisation meeting where it was clear that the recommendation had been implemented.

Another scrutiny raised as an issue the similarity between one examination question and an example in a recently-published textbook written by two of the examiners. The awarding body took action by removing one of the examiners from his position and deciding not to renew the contract of the other when it expired the following year.

In addition to the individual scrutinies, with their action plans and follow-up, QCA analyses the recommendations made across all scrutiny reports for each awarding body to identify any common themes or trends. These generally relate to procedures rather than being subject-specific. Anything subject-specific is followed up on syllabus-by-syllabus basis although each report requires the awarding body to review its whole provision in the light of the recommendations.

Code monitoring

In addition to in-depth scrutinies of individual examinations, QCA carries out a broader sampling of awarding body activity through its code monitoring work. Trained observers attend meetings held by the awarding bodies at various stages of the examining process during each examination series and report on the extent to which the meeting has adhered to the procedures specified by the code of practice.

Reports from observers are analysed to identify any general patterns. QCA then produces a confidential report for each awarding body identifying general issues

requiring attention and providing detailed comment on areas of concern relating to particular qualifications.

In 2004, this monitoring established that many awarding committees did not always consider the implications of their decisions about grade boundaries for a course unit for the aggregate grade outcomes for the whole subject.

Monitoring standards

A key issue in public discussion of the examination system is whether 'standards' are being maintained. QCA conducts a rolling programme of reviews across time, subjects, awarding bodies and qualifications. The overall aim of these reviews is to determine if any action is needed to safeguard examination standards.

Most comparability activities investigate both 'examination standards' – the expectations of students defined in the curriculum and expressed in the examinations – and 'performance standards' – the actual levels of learning that students achieve and the distribution of grades awarded in recognition.

Reviews of subject standards over time

Since 1997, QCA has reviewed standards over time in around six subjects per year. The reviews generally involve an investigation of the standards of work within a subject, across the awarding bodies, over a twenty-year period. Within the last two years, QCA has reviewed several subjects for the second time. These second reviews take the first review as their starting point and typically involve a shorter time period of five years or less.

The programme has been structured to cover every major GCSE and A level subject and 31 reviews have been completed and published to date. We reviewed the reports on French, Geography, History, Mathematics, Physics and Science listed in Annex 3.

Reviewers are asked to analyse syllabuses and their associated assessment materials systematically using criteria specific to each subject. Candidate work at key grade boundaries is compared over time and across awarding bodies and reviewers are asked to produce a description of performance typical at each grade boundary reviewed. In all cases, judgements are made by several reviewers. The reviews to date have involved around 400 independent specialists reviewing around 9000 examination scripts, drawn from all the awarding bodies, together with their associated syllabuses, question papers and mark schemes.

A standards review aims to answer three main questions:

- What changes have there been to the examination over the period of review?
- What have been the effects of those changes on the requirements made of candidates?
- Is there any evidence of grade drift?

Candidate work used in standards reviews is drawn from a national archive of examination materials, maintained by QCA. Each year awarding bodies send syllabuses and their assessment materials, as well as a sample of candidate work, in a range of specified subjects to QCA to be catalogued and stored for use in monitoring work. The national archive was set up in 1997 to ensure that a sound evidence base would be available to support investigations into standards over time and across awarding bodies in GCSEs and A levels.

As the reports of these reviews make clear, comparing examination standards over time is a complex task, heavily dependent on the evidence available and the ability of reviewers to make valid judgements on it. When considering the findings of standards reviews, several limitations need to be kept in mind:

Changes in syllabus and examination content
Syllabuses and examination papers may have changed significantly, particularly over a twenty-year period. Fundamental changes make it difficult for reviewers to make valid judgements about relative standards because they are not comparing like with like.

Individual opinion
Each individual places different values on each part of a subject. Agreed definitions of standards and frameworks show reviewers the standards they should work to, but it is difficult for them to avoid applying their own values. This can lead to differences in opinion about the same syllabus or script.

Lack of evidence
Reviewers do not always have all the evidence they need, particularly from earlier years. For example, whilst syllabuses and question papers are usually available for all years, mark schemes are sometimes missing and only limited candidate work is available from the earlier years. Reviewers regularly comment on the difficulty of appraising incomplete candidate work. The creation of the national archive in 1997 ensures that both examination materials and samples of candidates' work are available for comparisons commencing from that year.

These reviews of standards over time can provide only tentative conclusions on whether either 'examination standards' or 'performance standards' are being maintained. Baker et al. (2002) reached this clear conclusion for long-term reviews on the database available before 1997. Even with this new database, the reviews will inevitably be restricted to tentative conclusions for the first of the three reasons given above.

What the reviews will usefully provide are accounts of the often radical changes that take place in the content and assessment of many subjects over time. In history, for example, the 20-year review documents the shift from examinations focused almost entirely on knowledge to examinations seeking evidence of skills in interpretation and evaluation. (It needs to be recognized, however, that teachers and schools and colleges often see the task of examination preparation as seeking to ensure that students are able to deal with questions requiring interpretation and evaluation as questions requiring only knowledge.)

Reviews of standards across subjects

From time to time, concerns are raised about whether the standards required to achieve success are the same across different subjects.

One approach to answering this question is quantitative. It can involve investigation of the numbers of candidates achieving at different grade levels in different subjects but that comparison can be confounded by differences in the nature of candidates in different subjects. It can involve a comparison of candidates' prior achievement levels (e.g. GCSE for analyses of GCE results) but that depends on the comparability across subjects of the prior achievements. It can involve comparisons of performances in a given subject in other subjects taken. All of the Australian systems use this last kind of comparison and actually adjust marks in each subject on the basis of whether the results of its candidates in their other subjects show them, overall, to be above or below average. There are strong assumptions about the comparability of all subjects required for that approach but it would be difficult to apply in the case of GCE A levels where students typically take only two other subjects than the one being examined. In Australia, they take four or five additional subjects.

The other approach is more qualitative and involves judgements of the comparability of examination demands and assessments of student performances across subjects. The method depends on finding experts who are sufficiently qualified to make comparisons across at least related subjects. This method was recommended by Baker et al. (2002) and initial pilot work suggests that it is workable.

QCA has initiated four comparative studies: i) history and geography; ii) English, history and media studies; iii) biology, psychology and sociology; and iv) the sciences. Reports on the first three studies should be published by early 2005.

Reviews of standards across awarding bodies

From time to time, QCA has concerns that grade standards in a subject might not be comparable across the different awarding bodies. When such concerns are raised, an initial analysis of statistical evidence is conducted to clarify the nature of the problem. If concerns persist, detailed comparisons are made of examination demands and grading standards applied in the syllabuses from the different awarding bodies.

For these analyses, awarding bodies provide syllabuses, question papers and mark schemes, as well as examples of candidates' work at specified grade boundaries. Teams of reviewers, comprising senior examiners from the relevant awarding bodies and independent experts, consider both the materials and scripts to see if there are any noticeable differences between the examinations. Review of coursework does not normally form part of these exercises, although coursework requirements are considered in evaluating the demand of the syllabuses.

A confidential report on the findings is sent to the relevant awarding bodies. This includes recommendations of action required by the awarding bodies in

order to rectify any differences in grading standard. QCA checks that appropriate action has been taken by including the relevant examinations in future quality control activities, such as the scrutiny programme or code monitoring.

A recent example of this kind of review involved a language other than English for which the relevant language community had raised concerns about the comparability of courses (GCSE) from two awarding bodies. The investigation suggested that the concerns were well-founded. QCA then worked with the two awarding bodies to establish a common understanding of the appropriate standard and agreement on the actions that each awarding body then took to bring their standards into line.

Cross-qualification comparability studies

Concerns may also be raised about the comparability of the standards of English qualifications and those of other qualifications with a similar level of recognition. Cross-qualification comparability studies involve comparisons of standards in the same subject and year between two different qualifications. This type of study is particularly challenging as it often involves comparing qualifications with different grading systems as well as different syllabuses. QCA has recently published a report on a study to compare standards in English, Mathematics, History and Chemistry in GCE A level and the International Baccalaureate.

Follow up to report on maintaining A level standards

An earlier review of standards of assessment in GCE A level examinations (Baker et al., 2002) recommended a range of research activities that the QCA should undertake itself or commission. The report identified a particular set of research topics which would "improve QCA and awarding body practices". A summary of progress against these recommendations is set out below.

Work undertaken

Comparability of examination questions

It has not proved feasible to trial sets of questions and marking schemes due to the reluctance of schools and colleges to participate in this activity because their students are already taking a significant number of examinations. An alternative approach is proposed in which both teachers and students will be asked to comment on approach and question difficulty.

Qualitative analysis of the content and cognitive requirements of papers and scripts in different subjects as described earlier in the discussion of reviews of standards across subjects (p. 20).

Psychometric properties of examination papers

Baker et al. (2002) recommended that there be systematic analysis of the psychometric properties of examinations papers to provide examiners with more thorough and sophisticated analyses of the performance of the examinations that they had set. This requires analysis of performance data at question level and these data have not been systematically collected by the awarding bodies.

The introduction of electronic distribution to examiners of students' responses at the question level (e-marking), described earlier (p. 8), yields data at the question level and will permit this work to start. Data from summer 2004 is now available for some A level examinations and QCA will receive reports on the properties of the examinations involved.

Quality of marking

Various proposals were made on monitoring of marker performance to ascertain whether date and time of marking, and background of markers, affects their marking. Some of these studies have been done. In general, there are no simple relationships between quality of marking and various background characteristics of markers. The introduction of electronic distribution of students' answers to individual questions, and marking of these on screen with immediate submission of results, opens up possibilities for real-time monitoring of markers and this possibility is being worked on by the awarding bodies.

Double marking in Biology, with the second marker not knowing the marks awarded by the first marker, is being carried out to determine whether this would significantly reduce errors of measurement. If the approach is successful,

a similar study will be carried out with English, which is a subject particularly difficult to grade consistently.

Use of uniform marks versus grades

Baker et al. (2002) suggested that QCA investigate whether universities would be willing to use results on the uniform mark scale (see p. 12) rather than the A-E, U letter grades, to minimise misclassification at grade boundaries and to assist in selecting among candidates with all A grades. This debate has, if anything, intensified since 2002 as some universities have increasingly complained that, with more students with A-grades applying for places than there are places available, the A-grade no longer provides sufficient information to help with selection.

QCA has engaged in public and private discussions about how to deal with the problem but appears not to have had the authority to forge a solution. QCA held a technical seminar to move this discussion forward and the QCA Chief Executive canvassed publicly the options of using results on the uniform mark scale or subdividing the A-grade into narrower categories. The awarding bodies seem to have waited for QCA to act. The Universities & Colleges Admissions Service (UCAS) was hesitant because of the additional information with which universities would have to deal but, in summer 2004, it began consulting the universities and the awarding bodies to determine their view on the use of marks in addition to grades.

A further potential complication was the imminent report from the Tomlinson Committee. It has recently reported and has proposed splitting the A-grade for some A level courses into A++, A+ and A to provide additional differentiation among students (Working Group on 14-19 Reform, 2004, paragraphs 177-178).

It seems to us unfortunate that the QCA was not able to deal more decisively with this issue since the utility, if not the integrity, of A level results was at stake. It may also have strengthened two developments of alternatives to GCE A levels ahead of a fuller discussion of their utility.

One is that the University of Cambridge Local Examinations Syndicate, the parent company of OCR, has developed experimental *Thinking Skills Assessment* tests (http://tsa.ucles.org.uk/index.html) that some Cambridge Colleges have used in 2004 as an element in the information that they collect on candidates. The other is that the Admissions to Higher Education Review (Schwartz, 2004) has canvassed the possibility of using generic or subject-specific tests of aptitude.

The US College Board introduced the *Scholastic Aptitude Test* (SAT) in 1926 as a test of generic skills for use in university admissions. The SAT was introduced and used in the US in very different circumstances from those in England. In the US, there are no syllabus-based, public examinations of students' learning but only school-based assessments that were not comparable. The College Board later developed a series of 22 subject tests that provided measures of achievement, though still without direct connection to any particular curriculum. In 1990, as something of a retreat from the claim about measuring aptitude, the

name of the SAT was changed to *Scholastic Assessment Test*. In 1994, it became simply SAT. The original SAT, with its verbal and mathematical reasoning subtests, became the SAT I and the subject tests became the SAT II. In 2001, Richard Atkinson, President of the University of California, proposed that the University of California cease to use the SAT I precisely because it was a measure of 'aptitude' and not a measure of students' actual learning of the type that is provided by public examinations, like the GCE A levels (Zwick, 2004). The SAT has since been restructured in response to Atkinson's criticisms. Among other changes, items using verbal analogies (ones to which Atkinson took particular exception) have been eliminated and an essay and a mathematics material reflecting university preparatory mathematics courses have been added.

Validity of A level predictions

Baker et al. (2002) made a number of proposals for investigating the predictive validity of A level results in relation to performance in a range of university courses. No progress has been made in this area (and it is a moot question as to whether QCA considers this to be part of its remit). A final proposal, to compare the demands of A level with those of similar examinations in another country has not been carried out, although a study comparing the International Baccalaureate with A levels in Mathematics, English, History and Chemistry has been done, as mentioned earlier (p 21).

Summary

The e-marking project has particular potential to provide psychometric data on individual questions and to monitor the consistency/quality of markers. A comprehensive strategy needs to be developed, with consideration of the levels/types of data available, when they will be analysed and how the information will feed back into awarding body practice.

As for the other research suggestions, only limited progress has been made. The question we pose here is whether QCA has the resources (or indeed the remit) to carry out such a programme of research itself. It would seem more appropriate for QCA to set an agenda/strategic plan for research projects to improve QCA and awarding body practices which it then commissions from the awarding bodies and other agencies.

Risk analysis and management

Whether or not the QCA had a robust risk analysis programme in place prior to summer 2002 is not clear. What is clear is that the two main delivery risks collided, leading to results in which some stakeholders lacked confidence and a consequent major public criticism of the examination system.

We consider the two main delivery risks to be nested and to contribute to the overall responsibility of QCA, as the examination regulator, to maintain standards and public confidence in those standards.

The premier risk is *overload*. The introduction of AS-level in 2000 markedly increased the number of papers which have to be prepared and marked. It should be noted however, that the shift towards modular structures had already, in any case, started to increase the load on the system. The combination of the introduction of Curriculum 2000 and modularisation together are judged to have increased the traffic for schools/centres by 2½. The 2001 Baker Review into standards in A-level commented on the 'reported difficulties in obtaining a sufficient number of qualified markers' (Baker et al., 2002, p. 15) which had been exacerbated by the introduction of the AS level exam.

The function of risk analysis is to make individuals and organisations analyse risk(y) scenarios and implement mitigating action. QCA has been publicly recognising overload as an issue since the initial reviews of Curriculum 2000 in 2001 but the problem remains severe. In 2003, there were 26,874,636 papers ("transactions") across GCSE, AS, A2, GNVQ, VCE and AEA examinations with 60,404 examiners involved. The number of secondary school teachers in England is around 200,000. It is clear that the current exam system consistently operates near capacity, while being required to deliver extremely high quality results.

In the immediate aftermath of the problems with some A level examination results in 2002, the new Chief Executive of QCA established and chaired an Examinations Taskforce charged, among other things, "to assess risk to effective delivery of the examinations and take such action as is necessary to avoid it" (Boston, 2002, p. 10).

The setting up of the National Assessment Agency (NAA) is a subsequent, significant step in the process of risk management for QCA. Its role is to modernise the delivery of National Curriculum Assessment and the management of the examination system, using a combination of business process review and logistics analysis. It is through the NAA that the issues of 'overload' on the examination system and on schools and colleges (with more papers available and more opportunities for re-sits, finding invigilating staff and scheduling rooms for exam periods has become increasingly difficult) are being addressed.

Any significant reduction in the number of papers taken will require fairly major policy changes to, for example, the modular structure of courses, the balance of 'coursework' assessment versus external examination, and Curriculum 2000; all

issues which fall under the remit of the Tomlinson Committee. The issue here for QCA is whether political will and public opinion will allow a reduction in the number of examination papers taken.

A contributing factor to 'overload' is the supply and monitoring of trained markers. This is an issue which has to be addressed in the present while future policy on examinations is developed. NAA is addressing this as part of its remit; proposals include: recruitment campaigns, increasing the rate of pay, suggesting that each school 'supply' at least as much marking capacity as it 'demands' in examination uptake. Progress here is likely to depend to an extent on the response of teacher unions and professional bodies.

The second risk, related to overload, is *reliability*. Reliability in this setting relates essentially to consistency. Consistency in the examination is achieved by having common unseen papers taken in the same conditions. Consistency in marking is achieved by training markers, monitoring their performance and moderating results. If there is a problem in supply of markers, reliability is likely to be at risk as training may be hurried or inadequate; the marking itself may be sub-optimal as markers go too fast because of their load, and continue when tired. Under current arrangements, only limited ongoing monitoring of marker performance is possible and that exacerbates the risks resulting from poor marker performance. The NAA is addressing these issues: proposals include the professionalisation of markers (the National Institute of Assessment has been announced on the model of a College of Examiners) and shifting marking online which will allow real-time checking of marker performance. In summer 2004 Edexcel marked 25% of their scripts (GCSE, AS and A2) online; this speeded up the process [results were ready 7-10 days earlier than scripts marked traditionally] and allowed for closer monitoring of markers' performance.

One common proposal for securing standards is the double (or multi) marking of scripts. Whilst advantageous to quality assurance this would increase the burden on markers and, though important as a medium-term goal, is probably not achievable in the immediate future.

Other risks inherent in the delivery of the system result from the physical movement of scripts around the country and the number of communications required between schools and colleges and awarding bodies, which can lead to loss of data, delay and confusion. These are also being addressed by the NAA modernisation programme. It is anticipated that more use of electronic communication and electronic movement of scripts will mitigate these risks. A move to online activity will, of course, lead to other risks commonly associated with IT-based systems, of which NAA will be well aware.

Another risk is related to security: currently examination papers are delivered to schools and colleges three weeks before the exams start and have to be kept secure. Some minor 'security alerts' happen each year. This is a particularly tricky risk to mitigate: although all examination schools and colleges are visited and inspected routinely by the awarding bodies, QCA is not in a position to control this risk. The likelihood of occurrence is hard to estimate but the impact could be severe; the last resort is for the awarding bodies to produce a completely new examination paper if a security breach is uncovered before the

examination date. Awarding bodies take security very seriously and their ultimate sanction – to withdraw examination centre status – is one which schools and colleges similarly take very seriously.

Overall, we judge that the NAA will play a significant role in mitigating and managing the risks in the examination process thus supporting QCA in its role in regulation of examination procedures and in maintenance of standards.

Conclusions

The qualifications system regulated by QCA is sophisticated and complex, covering curriculum, accreditation of courses and examination of students. Much is expected of it – perhaps too much – so it often operates under extreme pressure of tight timelines and intense scrutiny.

QCA has a clear sense of its prime purpose, and this gives a sharp focus to its work. Its Annual Operating Plan for 2004-05 states:

> The mission of the organisation is thus to ensure, for all learners, fairness, access and equity, in relation to curriculum, qualifications and assessment.

Our remit is to focus on only one small, but important, part of the system and QCA's regulatory role within it. That is on 'examination standards' and we have chosen to focus specifically on GCE A levels.

In this report, we have described in some detail the examinations procedures operated in England under the regulatory authority of QCA but we note again that similar structures operate in Wales and Northern Ireland. We did this to make clear the care with which the entire examination enterprise is conducted. We have also described in some detail the way in which QCA exercises its regulatory role in respect of examinations to make clear the way in which it interacts with the awarding bodies. We have, in addition, given some illustrative examples of changes in procedures that have flowed from QCA's quality assurance and control procedures.

We turn now to our more specific conclusions and recommendations for change.

On the system

- No examination system at the school or other level is so tightly or carefully managed.

  Most countries have self-regulated government departments or agencies conducting examinations. In England, there are three large and highly professional, independent awarding bodies operating within a tightly regulated environment with many checks and balances,.

  This system is not without its problems, of course, some of them unintended, or inadequately anticipated, consequences of related decisions. The most notable example is the essentially simultaneous introduction of a modular course structure and AS courses, both of which increased enrolments and examinations in a system already under stress.

On standards

Discussions about whether standards are being maintained are made complex by differences in usage of the term 'standards'. We believe that it is important and helpful to distinguish expectations of levels of learning to be achieved,

examination standards, and the actual levels of learning achieved by students, *performance standards*.

Questions about whether standards are being maintained over time, across awarding bodies, across subjects and in comparison with other qualifications need to be addressed with respect to both examination standards and performance standards.

We conclude that:

- Strategies for maintaining comparable examination standards across awarding bodies are adequate to the task.

  The *Code of Practice* and the collaborative way in which it is developed between the regulators and the awarding bodies generates and keeps up-to-date and clear and workable set of principles and practices to which all parties can be held.

  If there were only a single awarding body, there would, on the face of it, be no risk of differences in examination standards in a single subject because there would be only one subject but, even within a single awarding body's programme there are options within subjects that render them multiple rather than single subjects in terms of curriculum and assessment.

- Strategies for maintaining comparable examination standards across time do as well as possible, but there are unrealistic expectations still in play.

  The cycle of subject reviews that QCA has initiated is building reasonable evidence of the extent to which standards are maintained in the relatively short-term of five years or so. The new national archive of curricula, examinations, mark schemes and marked student examination papers is a good support for this process.

  Over the longer term, it makes little sense in many subjects to ask whether examination standards have been maintained since the subjects themselves have changed so much. The prime value of a review with a long time horizon, such as the 20-30 reviews that QCA has conducted, is that it documents the nature and extent of changes in the curriculum.

- Strategies for determining whether comparable performance standards are being maintained are difficult to strengthen in the environment of public examinations.

  The *Code of Practice* specifies the manner in which awarding bodies are to use examiners' professional judgements of the quality of students' work in relation to the examination standards and evidence about the impact of their provisional judgements on the distribution of key grades (A, and E or better) in determining where to set ground boundaries in a particular year.

Judgements against the examination standards are *standard (or criterion) referenced*. Judgements that seek to keep grade distributions comparable from year to year are *cohort (or norm) referenced*. While the *Code of Practice* gives primacy to standard-referenced judgements, cohort-referenced judgements have also come into play before a final decision is made.

Other systems resolve this inherent conflict by abandoning one or other of the judgements. Many public examination systems abandon standard-referencing and essentially maintain a more-or-less cohort-referenced system by holding grade distributions roughly constant. This is true of France, Germany and South Korea, for example. In the case of end-of-secondary school assessments, when distributions are held constant while participation rates grow substantially, there can be no claim that performance standards are being maintained. The results awarded are simply not intended to reflect a constant standard.

Some examinations can provide consistent standard-referencing over time but they require a level of control over content (typically with some confidential assessment material repeated at least once) to provide the technical capacity to link between successive examination the scales on which results are reported. With examination standards held constant like this, any marked change in the cohort of candidates over time will be reflected in marked changes in the distribution of results (marks and grades). In the case of public examinations, two points need to be taken into account. The first is that the obligation to make examination papers and mark schemes public makes impossible the repeated use of material to link the scales of results psychometrically over time. The second is that, if participation rates were to grow rapidly as they have done in England over the last 20-30 years and as they might do even more quickly in the immediate future, holding examination standards rigidly constant and delivering high results (high marks and A grades) to a diminishing proportion of candidates would be unreasonable and, presumably, publicly unacceptable. No country has done it.

Some improvement in performance standards, as measured by examinations, is due to improvements in examinations. Strenuous efforts are made to compare performance levels over time, but the examinations have themselves changed notably towards a model of open disclosure to both students and teachers. The results of these necessary comparative studies are therefore often somewhat tentative.

- <u>No examination system has found an adequate way to determine whether standards are constant across subjects.</u>

  Neither QCA nor the awarding bodies have any strategy for determining whether examination or performance standards are constant across subjects. The only examination systems that we know of which seek to achieve this kind of comparability are those in the Australian States and Territories. They achieve this by making the rather heroic assumption that all examinations are measuring essentially a common dimension and

then express all results on a common scale. We described this approach briefly in the section ‘Reviews of standards across subjects’ (p. 20) where we also explained why it could not be applied to A level examinations even if it were thought desirable.

- The English public and its media expect too much of the public examination system at school level.

  Much of the public discussion of examination results in England is based on the assumption that results are standard-referenced with a degree of precision that cannot be delivered. It would help if lower expectations were set, not asking if performance standards are rising or falling but only asking if the examinations are making reasonable and appropriate demands of students and if the results work for the key purposes for which they are intended.

  Among those key purposes is the selection function that they must serve for universities and employers. If debate about whether performance standards were rising or falling had not been so unhelpfully dominant in recent years, the failure of A level results to serve adequately the selection function for universities might have been addressed much sooner and by the suppliers not the users. Results in A levels at the top end need to be more discriminating than the current A/B grade distinction. On whether this should be achieved by providing marks on the uniform mark scale or by providing grades for units of work within a subject, we express no view. We recommend, however, that QCA work with the awarding bodies, in consultation with universities, to develop and implement a solution.

On quality control

Quality assurance rests in the hands of both the individual awarding bodies and the regulatory authorities. Other countries typically have self-regulating bodies that are either parts of government departments or government instrumentalities. The US is different in having a few large not-for-profit private companies provide non-syllabus-based assessments of general and specific skills which supplement school-based assessments of student learning required by specific school-based subject curricula.

We conclude that:

- The awarding bodies have broadly consistent and well-regulated systems for setting question papers, managing marking and awarding grades.

  These rely heavily and properly on the professional input of practising teachers, and there are concerns about the continued supply of good quality examiners. The awarding bodies currently operate robust methods to ensure high standards of marking. The newly-established National Assessment Agency (NAA) is actively pursuing strategies to maintain a suitable pool of examiners, but these are at the mercy of other factors beyond their control.

- QCA has robust systems in place to monitor and regulate the work of the awarding bodies.

  There are very clear procedures which awarding bodies must follow. They are meticulously implemented and regularly reviewed with an eye constantly to improving the examinations system.

  There are also comprehensive reviews carried out by QCA on a well-planned cycle to monitor performance and standards.

- Implementation of electronic distribution of student examination answers at the question level will improve control of marking and monitoring of markers.

  The capacity to divide student papers into separate questions for marking brings considerable advantages which we have discussed earlier in the section on 'Marking completed examination papers' (p. 8). It will be important that these not be oversold and also that they not be misunderstood. The process should not be described as 'electronic marking' or 'e-marking', for example, since the marking is still done by human markers. Baker et al. (2002) did refer to developments in computer marking, even of essays, but that is not being contemplated for public examinations in the UK.

On the research agenda

The Baker Report (Baker et al. 2002) identified a set of research studies that improve understanding of the examination process in England and that could potentially improve the policies and practices of QCA and the awarding bodies.

We conclude that:

- Limited progress has been made in addressing the research questions proposed by Baker et al. (2002) and we question whether sufficient resources have been allocated to this work.

  We note that Baker et al. (2002) proposed that QCA 'adopt a research-oriented stance' and offered its list of research questions as only 'for consideration'. Neither Baker and her colleagues nor we would be disappointed if some of the questions were set aside after active consideration of their value and the feasibility of addressing them. Our concern is that they have not been sufficiently addressed because inadequate resources have been provided for the purpose.

- Consideration should be given to commissioning awarding bodies and other agencies to carry out some research tasks within a framework of a coherent and phased research programme.

  In observing that insufficient resources appear to have been provided for research, we do not necessarily expect that additional resources be allocated exclusively within QCA. We expect QCA to develop a strategic

plan of research to enhance quality, validity and reliability, to undertake some of the work itself but to commission some of it to the awarding bodies and some to other research agencies.

On risk management

The major risk of overload has been recognised and steps taken to mitigate it. The related risk of (poor) reliability in marking is being addressed through the NAA modernisation agenda, with electronic distribution of students' completed scripts playing a key role.

We conclude that:

- QCA has undertaken an exhaustive analysis of risks and the establishment of the NAA to handle delivery issues is a welcome development.

  The delivery of the examinations has been subject in recent years to increasing levels of risk. While overload and reliability are the two greatest areas of risk, the physical vulnerabilities in storage of question papers in schools and colleges and in movement of students' completed papers are also of great concern.

On communications and public relations

QCA has a responsibility to maintain public confidence in the examination system and not only by diligently working to ensure quality. QCA possesses good quality data and procedures. The eighth recommendation of Baker et al. (2002) was that "QCA should expand its communications programme to help the public and the profession understand the benefits and limits of its testing programmes and of any modifications being introduced".

In 2004, QCA has acted strongly on this recommendation. To provide an accessible public description of how the system works, it has produced the *A Level Guide* and promoted it through the media and distributed it to schools, In June-August 2004, QCA organised briefings of various stakeholder groups, in the business community as well as in education, Numerous articles by the Chief Executive Office of QCA have been published in the press. QCA established an online Q&A service, and a dedicated helpline and it created the character of 'Dr. A Level' under which nom-de-plume a retired expert has been engaged to answer questions on all aspects of how the system works, on radio, and TV and in newspapers.

We conclude that:

- QCA has substantially improved its performance in communicating with professionals and the public about its procedures. It should continue to be proactive in public discussion, based on good quality data and rigorous evidence of the efficacy of the procedures that it and the awarding bodies use. To the extent possible, QCA's voice should be

distinguished from that of the government of the day to preserve and strengthen its role in the discussion of examinations policy and practice.

This will require strengthening of the culture-shift within the organisation that is reshaping its terms of engagement with the public in general. Overall, QCA and the awarding bodies have a good story to tell, and QCA need not be reticent in articulating it.

It is sometimes unclear where responsibility lies in pushing on with an important public agenda, particularly between QCA and the awarding bodies. We believe that the regulator should take the lead when the issues relate to the system as a whole.

There may also need to be something of a culture-shift in government, with QCA rather than Ministers accepting the responsibility for engaging in much of the public discussion about examinations.

References

Baker, E., McGaw, B. and Sutherland, S. (2002), *Maintaining GCE A level standards: The findings of an Independent Panel of Experts*. London: Qualifications and Curriculum Authority. (http://www.internationalpanel.org.uk)

Bennett, J. (2001), Standards-setting and the NSW Higher School Certificate. Sydney: New South Wales Board of Studies (mimeo). http://www.boardofstudies.nsw.edu.au/manuals/pdf_doc/bennett.pdf

Boston, K. (2002). The new agenda for the Qualifications and Curriculum Authority. Address to the QCA Annual Conference 2002.

McGaw, B. (1997), *Shaping their future: Recommendations for reform of the Higher School Certificate*. Sydney: Department of Training and Education Co-ordination.

OECD, (2001), *Knowledge and skills for life: First results from PISA 2000.* Paris, OECD.

Stobart, G. (2000), Maintaining and monitoring standards over time: Discussion paper commissioned by QCA. London: Qualifications and Curriculum Authority.

Tomlinson, M. (2002), *Inquiry into A Level Standards: Final Report*, London: Department for Education and Skills. (http://www.dfes.gov.uk/alevelsinquiry)

Zwick, R. (Ed.), (2004), *Rethinking the SAT: The future of standardized testing in university admissions,* New York: RoutledgeFalmer.

Annex 1: Committee's membership and brief

The Independent Committee on Examination Standards was established by the Board of the Qualifications and Curriculum Authority (QCA) on the basis of a recommendation in the final report of Tomlinson's inquiry into A level standards in December 2002.

> I now recommend that QCA should establish an independent committee whose role would be to review and, if necessary, advise QCA publicly on whether or not standards are being maintained – advising on a limited number of subjects each year - using all the available evidence including subject syllabuses, students' work, mark schemes and question papers. The group should also be able to review and verify other aspects of QCA's regulatory work, as requested by the QCA board. This committee will help provide reassurance that standards are being kept continuously under review and that, where necessary, action will be identified and taken to safeguard standards over time (Tomlinson, 2002, p.10).

The Committee was appointed in late 2003 and comprises:

- Dr Barry McGaw, Director for Education, Organisation for Economic Co-operation and Development (OECD) – Chair;
- Professor Caroline Gipps, Deputy Vice-Chancellor, Kingston University;
- Mr Robert Godber, former Headteacher, Wath upon Dearne Comprehensive School, Rotherham.

The Independent Committee is expected to meet three times a year with its work to cover GCE A level, GCSE and GNVQ examinations and to report direct to the QCA Chief Executive. For this, its first report, the Committee has chosen to focus on GCE A levels.

The main work of the Committee is expected to involve:

- evaluating the judgements made by experts in QCA's rolling programme of reviews of standards over time and other comparability work to ensure that they are well-founded and fit for publication;

- overseeing any follow-up action to investigations of comparability and standards over time that is needed to align standards;

- commenting on methodologies used by QCA in its comparability work.

After an initial briefing meeting in November 2003, the Committee met to undertake the work for this report on 23-24 March, 3 September and 7-8 October 2004. Its main sources of direct information are set out in Annexes 2 and 3.

Annex 2: People interviewed by the Committee

Qualifications and Curriculum Authority

Dr Ken Boston	Chief Executive
Mr Angus Alton	Programme Leader, Comparability, QCA
Mr John Barwick	Programme Leader, Monitoring, QCA
Dr Jonathan Ford	then Director, Examinations and Testing, QCA
Mr Chris Jones	then Director, Curriculum and Assessment Policy, QCA
Mr Dennis Opposs	Head of Quality Assurance Programme, QCA now Director of Quality Assurance, QCA
Ms Pauline Sparkes	Programme Leader, Quality Improvement, QCA; now Programme Manager, Examinations Management, NAA
Mr Mick Walker	Head of Examination Series Management, QCA now Director, Examinations Management, NAA

Awarding bodies

Dr Mike Cresswell	Director-General, AQA
Mr John Kerr	Chief Executive, Edexcel Foundation
Dr Ron McLone	Director- General of Assessment, University of Cambridge Local Examinations Syndicate [OCR)

In addition, both Professor Gipps and Mr Godber attended separate meetings of A level examiners to observe the procedures employed. Dr McGaw had made similar visits during the work for an earlier report on the maintenance of A level standards (Baker, McGaw and Sutherland, 2002).

Annex 3: Other sources of information

Papers commissioned by QCA

Adams, R. (2000), Test equating at GCSE and GCE: A commentary.

Bell, J.F. (2000), Review of the use of Thurstone Pair methodology to monitor examination standards over time.

Brooks, G. (2000), Review of models for maintaining and/or monitoring survey-based reading standards over time.

Cresswell, M.J. & Baird, J. (2000), A review of models for maintaining and monitoring GCSE and GCE standards over time.

Newton, P. (2000), Maintaining standards over time in national curriculum English and science tests at Key Stage 2.

Stobart, G. (2000), Maintaining and monitoring standards over time: Discussion paper commissioned to review the other five and additional research evidence.

Other papers

Baird, J., Cresswell, M.J. & Newton, P. Would the *real* gold standard please step forward?

Newton, P. Linking standards across examinations: Contrasting definitions of grade equivalence.

Pollitt, A. & Elliott, G. Monitoring and investigating comparability: a proper role for human judgement.

School Curriculum and Assessment Authority (1996), *Standards in public examinations 1975 to 1995: A report on English, mathematics and chemistry examinations over time.*

Schwartz, S. (2004), *Fair admissions to higher education: Recommendations for good practice*. London: Department for Education and Skills. (http://www.admissions-review.org.uk)

Working Group on 14-19 Reform (Chair: M. Tomlinson) (2004) *14-19 Curriculum and Qualifications Reform: Final Report*. London: Department for Education and Skills.

Reports commissioned by QCA

Baker, E., McGaw, B. & Sutherland, S. (2002), Maintaining GCE A level standards. London: QCA.

QCA Secretariat (undated), Humanities comparability (pilot) study.

QCA documents

QCA, ACCAC & CEA (2000), *Arrangements for monitoring and reporting publicly on external qualifications.* London: QCA.

QCA, ACCAC & CEA (2004), *Code of practice 2004/5: GCSE, GCSE in vocational subjects, GCE, VCE, GNVQ and AEA.* London: QCA.

QCA (2004), *Guide for principal scrutineers: Scrutiny programme for GCE, GCSE and GCSE in vocational subjects.* London: QCA.

QCA (2003), *Producing self-assessment reports: Guidance for awarding bodies*. London: QCA.

QCA audit reports on the general qualifications operation of individual awarding bodies.

QCA monitoring of post-audit action plans of individual awarding bodies.

Three scrutiny reports.

QCA standards over time reports

French: GCSE (1996-2001).

French: GCE A level (1977-1997).

Geography: GCE A level (1980-2000).

Geography: GCSE (1996-2001).

History: GCSE (1977-1997).

Mathematics: GCSE (1975-1995).

Physics: GCSE (1997-2002) and GCE A level (1996-2001).

Science: GCSE double award (1995-2000).