MAINTAINING

GCE A LEVEL STANDARDS

The findings of an independent panel of experts

Baker, E., Sutherland, S., & McGaw, B. (2002). *Maintaining GCE A level standards: The findings of an independent panel of experts*. London: Qualifications and Curriculum Authority.

JANUARY 2002

Contents

An independent panel of advisers was invited by QCA to review the adequacy of the quality assurance systems that are designed to maintain GCE A level standards. The standards concerned refer to the demands of the specifications and their associated assessment arrangements, as well as to the levels of the performance required of candidates to gain particular grades.

The three-member panel comprised Professor Eva Baker (Chair), University of California Los Angeles and Co-Director of the United States Center for Research on Standards, Evaluation, and Student Testing (CRESST); Dr Barry McGaw, Deputy Director for Education at the Organisation for Economic Co-operation and Development (OECD); and Lord Sutherland of Houndwood, Principal and Vice-Chancellor, University of Edinburgh.

In connection with the GCE A level, the panel was asked to:

review QCA's quality assurance work, including the outcomes of operational monitoring and comparability activities;

review the individual and collective quality assurance and quality control arrangements of the unitary awarding bodies (AQA, Edexcel, OCR);

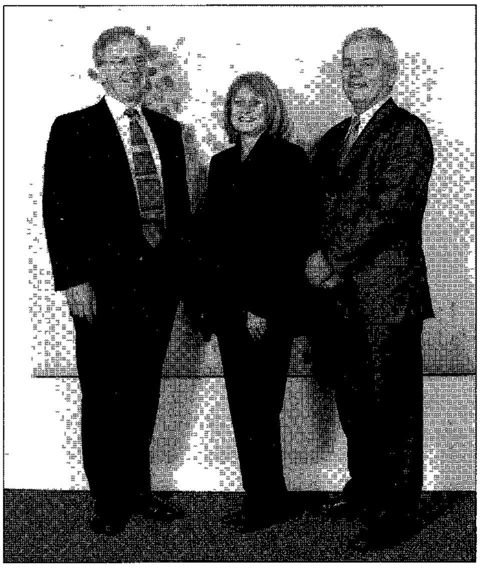
consider how the quality assurance and quality control arrangements operate in the context of specified subjects;

consider any other relevant evidence or opinion, including that relating to equivalent arrangements in other countries;

advise QCA on whether overall quality assurance arrangements match best international practice and how they might be strengthened;

publish its findings in late 2001.

The panel operated in a context where examination results achieved by candidates have improved year by year. This has raised questions in some minds about whether rising performance standards reflects declining demands in the examinations. The panel assessed the quality of examination systems based on the following criteria: accuracy, validity and fairness.



From left: Lord Sutherland of Houndwood, Professor Eva Baker, Dr Barry McGaw

Conclusions

Following numerous interviews with QCA staff, awarding body representatives, school personnel and university staff, and after reviewing documents and data, the panel has arrived at the following conclusions.

1. There is no scientific way to determine in retrospect whether standards have been maintained. Therefore, attention should be placed on ensuring the accuracy, validity and fairness of the system from now on.
2. It is our considered judgement that QCA is doing a commendable job to assure the quality of the A level examinations.
3. The awarding bodies are providing a good level of quality assurance within the very demanding timetable of the examination system.
4. Public policy and common perception may be in conflict on some points. In particular, the separate goals of improving schools and levels of student performance, and encouraging more students to pursue university study are at odds with the view that high A level grades must be limited to a very small proportion of candidates.
5. Some of the A level examination procedures might be improved but they would either add cost or time to the traditional process.

Key recommendations

1. QCA should adopt a proactive research stance in order to collect information that will bear on the technical quality evidence for A level examinations and standards. Three types of investigations are suggested: short-term studies of the comparability of examination questions and psychometric properties of examinations; ad hoc studies of the quality of marking and the use of uniform marks vs. grades; and long-term studies investigating the validity of the A level examinations in predicting university performance.
2. QCA should develop a strategic plan to ensure that methods can be employed to document the maintenance of standards in the future.
3. QCA should manage its role in a way that supports the examination process, exhibiting both transparency and accountability in its methods.
4. QCA should work to minimise unpredictability in requirements of the awarding bodies (and of schools and students). Imposition of new requirements with unreasonable timescales should be avoided.
5. QCA should not require changes that affect examination provision other than in a timely manner based on a clearly advertised schedule.
6. QCA should be aggressive in communicating with policy makers about the feasibility of their expectations, in particular when it is not possible for the system to deliver what is required on the timeline envisaged.
7. QCA should employ a convening function to air issues associated with standards in key areas, such as mathematics and science.
8. 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.

Introduction

An independent panel of advisers was invited by QCA to review the adequacy of the quality assurance systems that are designed to maintain GCE A level standards. The standards concerned refer to the demands of the specifications and their associated assessment arrangements, as well as to the levels of the performance required of candidates to gain particular grades.

The three-member panel comprised Professor Eva Baker (Chair), University of California Los Angeles and Co-Director of the United States Center for Research on Standards, Evaluation, and Student Testing (CRESST); Dr Barry McGaw, Deputy Director for Education at the Organisation for Economic Co-operation and Development (OECD); and Lord Sutherland of Houndwood, Principal and Vice-Chancellor, University of Edinburgh.

The remit

In connection with the GCE A level, the panel was asked to:

- review QCA's quality assurance work, including the outcomes of operational monitoring and comparability activities;
- review the individual and collective quality assurance and quality control arrangements of the unitary awarding bodies (AQA, Edexcel, OCR);
- consider how the quality assurance and quality control arrangements operate in the context of specified subjects;
- consider any other relevant evidence or opinion, including that relating to equivalent arrangements in other countries;
- advise QCA on whether overall quality assurance arrangements match best international practice and how they might be strengthened;
- publish its findings in late 2001.

The panel operated in a context where examination results achieved by candidates have improved year by year. This has raised questions in some minds about whether rising performance standards reflect declining demands in the examinations.

The process

The panel met three times in the period from February to October 2001 and, during these meetings, had discussions with various groups and individuals. A complete list of those interviewed is provided in appendix 1. They included QCA staff, senior staff from the awarding bodies, school personnel, higher education staff involved in admissions, students and Her Majesty's Chief Inspector of Schools in England. The panel was also provided with information and was supported in a variety of ways by staff of the three awarding bodies. Individual members of the panel visited the awarding bodies and were able to observe examiners' meetings about determining the awards of grades for 2001. They also met with a wide range of staff who were involved in developing and implementing A level examinations.

The structure of the report

This report is organised as follows. The core criteria used by the panel as a basis for its study are first described. Then each point in the remit is addressed, although there is some necessary overlap. In addition, the panel suggests some continuing research and evaluation activity that QCA should undertake to strengthen the maintenance of standards in the future.

Core criteria underlying the panel's deliberations

It became immediately clear that four complementary concerns were at the heart of the remit:

- the definition of quality in an examination system;
- the various means by which an erosion of standards could occur;
- the sufficiency of procedures for quality assurance;
- the feasibility of approaches to assure maintenance of standards.

All of these concerns lie at the core of the professional and public credibility of the A level examination system.

Before we address the specific questions in our remit, we believe it important first to declare explicitly the operating definition of quality that guided our work. There have been extensive efforts to describe desirable quality criteria for examinations. (See, for example, *Standards for Educational and Psychological Testing,* American Educational Research Association 1999; 1985.) We capture those that we used under three headings: accuracy, validity and fairness. We intend to avoid giving an extended lesson in measurement, but instead offer a brief summary of the meaning of each criterion because we think it will be useful to readers not only in interpreting our analysis and recommendations but also in considering future results and discussions of examinations.

Accuracy

Accuracy, in common language, refers to whether test results are correct. Components of accuracy include whether performance was correctly scored and whether the scores, if they are used to classify or grade students, do so appropriately. Correct scoring has a number of elements, including how the scoring guide is developed and the degree to which markers of student work adhere to the guide and make consistent judgements. Correct classification or grading depends on the precision and appropriateness of the cut-points between categories or grades established on the underlying mark scale, which in turn depends on the accuracy of the marks. The ultimate purpose is to assure that misclassification (giving a student an undeserved grade) is minimised.

It must be recognised that error is a reality in all measurement. Even measurements of physical characteristics such as weight, height and speed are fallible to some degree. The fallibility may be due to differences in the situation (for instance, the time of day at which a person is weighed) or it may be related to the precision of the instrument (for instance, weighing with a scale calibrated in half kilograms and so rounding to the nearest half kilogram).

Measurement errors are not necessarily deliberate mistakes. Even the most carefully developed examination, of the type reviewed in this report, will result in fluctuations in assessments among students and markers. The task in an examination system is to reduce error in a manner consistent with the goals of the system and the costs it can bear.

Validity

Validity is a technical term describing the degree to which the inferences drawn from test results are appropriate. In addition to a number of technical concerns to be raised, validity depends upon the purposes for which the results will be used. In A level examinations, one important purpose of the system is to provide a basis for selecting students for university admission. Another is the assessment and certification of students' level of mastery of subject matter.

Students' performances in the examinations are supposed to be based on their level of achieved expertise in subject matter learned through the instructed syllabuses. To this end, specifications are created from which courses with approved syllabuses are developed. The A level examinations in each area are supposed to represent the content taught in the course. Thus, the validity of an examination would be threatened if the examination questions did not represent the content to be learned, as expressed in the syllabus. Because schools have a choice of awarding bodies with attendant syllabuses, examinations and scoring procedures, it is possible that the validity of the examinations could vary across the bodies.[1]

Of particular interest in our review is the degree to which the validity of examinations has been maintained over time. It is also clear that the qualities of specifications that control the design of the syllabuses are central and must be the subject of broad agreement. We focused our validity attention, however, on the examination practices.

Another complication in a validity analysis is the potential for conflict in examination goals. On the one hand, the expectation is that students will be differentiated from one another, with marks or grades awarded to reflect such differences among them. In cases such as A levels, there is the further expectation that the highest grade will be awarded to a limited number of the very best students. On the other hand, there is an expectation that examinations will measure what students know and are able to do, with marks and grades reflecting the standard of their individual performances rather than their relative standing among candidates.

The differences among the assessment purposes of *selection* (eg choosing the best students for university admissions) and of *certification* (eg assigning high grades to those candidates who achieve the desired standard) have been the focus of persistent discussion in the literature on educational measurement.

Norm-referenced assessment (related to the certification purpose) focuses on comparing students with one another. Results follow a 'normal' or bell-shaped curve, consequently ensuring that very few students receive the highest marks. Criterion-referenced assessment (related to the certification purpose) attends to whether students have satisfied stipulated performance criteria and grants all students who achieve particular levels of performance the corresponding grades. No specific concern is held in criterion-referenced systems about the proportion of students who will receive the highest grades. One potential difficulty in interpreting A level findings is the desire to have both a norm-referenced and a criterion-referenced system. In public discussion of the distribution of awarded A level grades, two expectations appear to be held simultaneously:

1. that the allocation of high grades will be determined by a general *a priori* percentage of students thought to be an acceptable proportion in each category, harking back to the normal curve;

2. that the allocation of grades should depend exclusively on the merit of the performance of candidates.

A criterion-referenced system would produce shifts in the percentage of candidates receiving different grades to reflect concomitant shifts in the proportion of students achieving proficiency at each level. This approach is indifferent to the percentage of students in the age cohort taking the examinations.

Those adhering to a norm-referenced system may assume that a growing proportion of participation by the age group in examinations is inevitably accompanied by a drop in average performance and should result in declining numbers of students achieving higher grades. That assumption actually has two components. One is that the growth in participation comes exclusively from students who will perform worse than the smaller proportion taking the examination in the past. Its corollary is that it is not possible to raise absolute performance levels by reforms to educational practice. Both parts of the assumption are perhaps rooted in beliefs about the constancy of human performance and are at odds with experience in many areas of human endeavour, from performance in sports to productivity levels in business and industry. The assumption that there is a limit on high performance does not stand up to clear international evidence which shows that academic achievement is better in some countries than others, both in average results and in the proportion of students achieving high levels of performance. General research findings show that high-quality instruction, student effort and persistence pay off in academic achievement. We therefore recommend that the study of the validity of higher proportions of achievement be conducted. Clearly, a goal to improve the overall percentage of legitimate university entrants assumes that performance distributions of achievement can change.

The A levels present another interpretative challenge to validity, because the choice of which examinations to pursue is left to the candidate. These choices may be based on aspirations to study various university courses as well as to optimise admission to higher education. Thus, the composition of the cohort sitting for different subject examinations may shift at the same times as the overall numbers of those presenting themselves for A level study change.

Validity for examinations is not an all or nothing proposition, but rather a matter of interpreting available evidence to support conflicting claims. Compelling evidence may come from different sources and, in any case, never provides certainty. The best evidence has a strong scientific base and does not depend upon unsubstantiated opinion, no matter how strongly or persuasively it is expressed. Scientific investigations of validity can be undertaken. For example, performance on other comparable examinations might be used to determine whether validity inferences are warranted. Validity might also be investigated by having individuals who are known to be excellent in a particular subject area take the relevant examination to see how they perform, in contrast to those who are unlikely to be successful.

Fairness

Fairness ultimately addresses the question of whether students given the same quality of preparation and who have the same degree of motivation would be likely to perform similarly in the examinations in question. Fairness involves the extent to which the test administration and scoring practices are comparable across identifiable groups of students. Fairness also includes the degree to which students have equal access to information about the testing process itself, and whether the examination questions or scoring process have attributes that would disadvantage an identifiable subgroup. Our use of the term 'fairness' in this fashion is not intended to convey that the performances of particular subgroups should be more or less equal, although that use of the term is sometimes made. Differences in group performance may be due to differences in preparation, eg quality of teaching, access to support, motivation, as well as to any differences among the subgroups, such as English language proficiency. Technical analyses of fairness of examination systems should focus on elements in the examination process that can be systematically controlled.

Using the criteria

These criteria of accuracy, validity and fairness are core concepts in our analysis of the A level examinations and form the strands of our advice for monitoring quality. Together they should form the appropriate basis for the credibility of the examination system in the minds of students, professionals, policy makers and the public. Credibility, as we use it, simply means the confidence that people can place in the quality of the examination process and the legitimacy of its results.

Can we answer the question of whether standards have been maintained in retrospect?

The press and public have raised the question of whether standards are being maintained, with the implication that the A level examinations are not as difficult as they used to be and that performances receiving high grades are not as good as they used to be. They cite, in partial support of this thesis, the increasing proportions of students passing the examinations at a high level.

Some systems have given up on any attempt to address this question and restrict themselves to a limited normative approach to marking and grading. One common approach is simply to fix the percentages of students to be allocated each grade and to apply those percentages to every cohort of students – a clear norm-referenced strategy. Another approach makes that strategy even more explicit by reporting the students' percentile rank within the cohort, showing whether they are at the 99th percentile (top 1%), the 98th (top 2%), and so on of whatever is the relevant student cohort.

A variant of this approach has been introduced in Australia in an attempt to take some account of changes in the student cohort.[2]

To answer the question of comparability of standards and to make a legitimate interpretation of trend data over time, certain conditions would need to be met, including that:

- the specifications and syllabuses have remained constant;
- the examinations given are common over time or can be equated;
- there have been no changes in educational policy or practice intended to raise performance.

None of these conditions have been met as there have been numerous changes to the system. These changes include new syllabuses; the types and intensity of the professional development of teachers; increased openness of the examination process with more detailed reports to teachers from chief examiners and a change in what is regarded as an appropriate examination question (eg moving to analysis and away from memorisation), to name but a few. In the short term, the conditions to allow the interpretation of trend data might more readily be met and current practice assumes that they can. Meetings of examiners use sample scripts at key boundaries, such as A/B, from the previous year to check consistency of marking and grading over a short term. There are a number of contending reasons for successively improving performance which include:

- true improvement in student achievement;
- more widely diffused knowledge about what is likely to be tested;
- changes in the specifications and resulting syllabuses;
- easier examination questions;
- less stringent marking criteria;
- lax application of scoring criteria;
- changes in methods for establishing boundaries.

More than one of these elements may be present simultaneously. While in the subsequent discussion we consider these interpretations, it is clear that an unyielding focus on the issue of the past is of little value. We wish to direct QCA's attention to strategies and procedures to assure that standards are maintained in the future, so that contention on the point of standards maintenance can be reduced (although surely never eliminated).

Task 1: Review QCA's quality assurance work, including the outcomes of operational monitoring and comparability activities

In pursuit of this first task, the panel met with appropriate members of QCA staff, as indicated in appendix 1, to learn of their procedures for quality assurance. We also had repeated interactions in person and in writing with the awarding bodies. It is our judgement that the monitoring activities are appropriate for a regulatory agency, which must find a balance between reasonable oversight to protect the public and intrusion.

Although responsibility for regulating external qualifications in England lies with QCA, the awarding bodies also bear responsibility for quality assurance. Statutory regulation is used to safeguard the public interest where other mechanisms, including awarding bodies' own quality assurance systems and quality control arrangements, would not be sufficient. Regulation, through accreditation and monitoring, is based on accreditation criteria, including the common code of practice and qualification-specific codes of practice. These requirements are used in carrying out monitoring, drawing conclusions about aspects of an awarding body's work, and reporting. They also provide the basis for awarding bodies' self-assessment reports. A number of related procedures are used to support QCA's quality assurance. These include annual statistical reviews of results and detailed scrutiny of individual examinations, involving the syllabuses, question papers, mark schemes, scripts and so on. Visits are made to each awarding body in the autumn to discuss its examinations in detail. A confidential report is provided to the awarding body, containing recommendations for the examinations. The awarding body provides a response to the recommendations in order to satisfy QCA that appropriate action has been taken. In addition, QCA has appointed teams of Principal Scrutineers to assess the consistency of procedures and comparability of standards across awarding bodies in ten key A level subjects. The teams of Principal Scrutineers operate in: English, mathematics, biology, chemistry, physics, geography, history, French, German, and design and technology. Further information about the role of Principal Scrutineers can be obtained from QCA.

Perhaps of most relevance to the question of standards maintenance are the five-yearly reviews investigating both changes in examination demand and changes in standards of performance. Details of this process are available from QCA but, in general, a review involves careful analysis of the entire process, including scripts of candidates from different years at the defined boundaries to determine whether standards have changed.

Given the continuing concern expressed about certain subject areas, eg mathematics, the panel recommends that QCA seek the resources to carry out the reviews on a more systematic schedule and, for selected subjects of concern, with shorter intervals between them than five years.

We believe there should be greater open discussion among QCA, awarding bodies, schools and representatives of higher education courses (eg mathematicians) on the nature of the specifications designed by QCA. We understand this will be a cost burden to both QCA and the awarding bodies, but believe it to be an important step in enhancing the attention given to the issue of comparability of standards across years. The panel is reassured about QCA's ability to monitor comparability across awarding bodies.

The question has also been raised about comparability of standards across subjects. To this concern, we can point to the processes that the awarding bodies and QCA use to assure quality. We doubt, however, that there is a method to document comparability of standards (meaning performance) across different subject areas. For one thing, the disciplines themselves differ in terms of the types of prior knowledge, analytical skills and requirements for expression. Second, the various subjects draw to them students with very different interests and capacities.

Of education systems, the Australians, alone among systems as far as we know, do attempt to achieve comparability of assessments across different subjects for the purpose of university admissions though not for the results that appear on students' certificates of achievement. These systems base the distribution of marks in a subject on information about how select the cohort of students in the subject is. This is estimated by using the students' average results in all their other subjects. The marks in the subject in question are adjusted so that the students' mean and spread match those of their average result in their other subjects. As a consequence, some subject results are set higher than others. The process of adjustment is repeated using the now adjusted marks until further repetition makes no difference. The adjustment is based on the selectivity of the candidates, rather than any measure or judgement of the difficulty of the subject.[3] This system for equating results across subjects would be unlikely to work in England, even if it were thought desirable. With students typically taking only three subjects, it is unlikely that there would be sufficient overlap between subject combinations to estimate comparable averages.[4] The provision for students to take subjects with more than one awarding body would also reduce the capacity to estimate sensible means.[5]

We believe that attempts to look at the performance of students in different subjects to discern differences in standards have little value. Better, we believe, would be to analyse the syllabuses, examinations and marking schemes for different subject areas according to the task demands they require of the student. A systematic assessment, using a well-defined analytic scheme with common elements across scrutinies and reviews, would provide useful information to guide any necessary modifications.

QCA should provide a plan that covers its intended monitoring activities for a particular period, for instance annually. It should address special policy issues of interest, eg the AS examination introduction, as well as its plan for regular quality assurance and quality control. The plan should explicate clear goals to be achieved in the period. In executing its plan, QCA should engage the Joint Council of the awarding bodies to assure cooperation.

Task 2: Review the individual and collective quality assurance and quality control arrangements of the unitary awarding bodies

We interviewed representatives of the awarding bodies and queried them about the approaches they took to make sure their processes were worthy of the public trust. We also observed actual procedures during the awarding process in the early summer. In addition, we were able to review documents provided by the awarding bodies, including the code of practice. We were impressed by the high degree of professionalism demonstrated by the awarding bodies' representatives. We believe that the awarding bodies are providing a good level of quality assurance within the demands of the examination system.

The demands themselves, however, may be worthy of consideration. Because the interval between the administration of the examinations and the release of results is very short – around ten weeks or so – many procedures have been traditionally used in order to speed up the process and some, we believe, introduce threats to quality. We provide only a few examples for illustration. For instance, we understand that examination papers are sent directly from schools to particular markers. This procedure avoids collecting the examination scripts in a centralised area and redistributing them at random to a set of markers. The present procedure results in two realities:

- the marker knows that sets of scripts come from a single school and, despite recent instructions that schools should not forward the scripts in envelopes that identify the school, the marker can often tell from which school the scripts come because of postmarks or the school using its own stationery;
- the scores of these schools in a particular subject are wholly dependent upon the particular marker to whom the papers are assigned. Furthermore, only one judge marks each script. There is considerable effort in the initial stage of marking to ensure that markers deal similarly with papers, with the results of initial marking reviewed by senior examiners. Further checks are made at later stages in an attempt to maintain consistency but it is predominantly the case that judgements are made by only one marker without any external checking.

A second consequence of the time constraint is the unsupervised marking of scripts. Markers receive scripts and are given guidelines about how they should go about their tasks. However, it is possible that an individual marker might procrastinate and mark many scripts more rapidly as the time to return them approaches. More likely, however, is that reading many papers over time could cause fatigue. Fatigue has been shown to affect some examination markers by subtly modifying the scoring scheme they use. Approaches to controlling these effects typically require rating sessions to be conducted in centralised locations.

Thus, some of the potential questions that might be raised about the marking process could be resolved by procedures that cost more and take longer to complete. In the light of reported difficulties in obtaining a sufficient number of qualified markers,[6] these approaches may not be feasible.

Another consequence of a traditional practice is worthy of notice: the annual procedures used to develop questions. The awarding bodies have extensive experience in carefully developing examination questions and using experts to create the marking schemes for judging candidates' performances. The examination development process occurs sequentially. Draft questions are set by a chief examiner, who will have had considerable marking experience before setting questions for the first time, and the questions are then reviewed by an experienced panel. Marking schemes are subjected to similar review. The process, however, proceeds for one examination at a time, even though an awarding body might well have the examinations for two successive years ready ahead of time and so have questions in reserve. The questions necessarily change annually for security purposes. This procedure is different from one in which many years of questions might be developed contemporaneously and assigned at random to various years, thereby attempting to ensure that each year's examination should have an equal chance of having comparably difficult questions. The serial development of questions and their assignment to examinations make it very difficult to judge the comparability of questions and marking schemes from year to year. As a result, from what we could determine, an effort is made to adjust for any apparent shift in the difficulty of the examinations from year to year by making the distributions of grades comparable, with similar percentages of candidates receiving marks within the various grade ranges to those in prior years. There will always be error of measurement involved in setting grade boundaries and, in any case, the shift of a boundary up or down by a single mark can alter quite considerably the proportions of students in the grades on either side. As a result, the percentages can seldom be exactly the same between two years, so they may, in fact, inch up. For example, if the proportion of students receiving an A grade increased by only 0.5% annually, the overall impact would be a shift over ten years of 5%, for example from 10% to 15%, of students achieving an A grade. This growth would occur as a result of the attempt to maintain comparability between two years of performance, combined, perhaps, with a tendency to upward rather than downward movement in the face of uncertainty. The current practice is for examination panels to examine grade distributions for prior years in an attempt to resist any artificial drift upwards. In addition, panels have archive scripts at the key grade boundaries in order to make direct comparisons of current and past student work as the boundaries are set for the current year. As will be discussed in Task 4 (which examines international practices) there are alternative procedures that might be employed, but at a cost.

There are also inevitable questions about competition among, or the need for, three awarding bodies performing similar functions.[7] Even though the awarding bodies grew from different orientations, we believe they show reasonable consistency in their practices, a reality probably attributable to a range of factors including the code of practice, the Joint Council, and QCA policies and reviews. As schools have a choice of syllabus, they are given a certain amount of flexibility. Competition among them should have a positive result and encourage innovation, provided constraints on practice do not inhibit innovation. We believe their differences reside in different syllabuses and examinations rather than different assessment practices. In our observations we found that, for the most part, procedures were comparable. We noticed some differences in the ways in which statistical information was used in meetings of examiners, but felt that difference did not materially alter the outcomes. We believe it worthy of study to determine why a school chooses a particular awarding body, and whether subject matter emphasis, professional support, or some other variables account for the choices made.

Task 3: Consider how the quality assurance and quality control arrangements operate in the context of specified subjects

Our attention to specific subjects occurred necessarily as we considered the key components of the remit. We believe that QCA will need to give more attention to some subjects than others in its monitoring and quality assurance programme because its resources are finite. We recommend that QCA make the criteria for determining its monitoring programme explicit.

At this stage, we believe that mathematics is clearly an area that currently warrants attention for a number of reasons. University officials report that remedial work is required for courses for which mathematics is a prerequisite. At the same time, it is unclear whether the nature of the candidature in A level mathematics has altered much. The proportion of successful students is rising but there has been a reduction in the number of students taking the mathematics A level examinations, amounting to a decline of approximately 13% over the last ten years. The growth in the percentage of students receiving A grades and the decline in the numbers of students studying mathematics at A level, combine to yield a steady 10% of the 18-year-old population achieving A grades across the decade. There is also continuing evidence that students taking A level mathematics are among those most successful in the GCSE.

What is the cause of this apparent conflicting information? Some new policies would seem to work to improve performance. Increased numeracy expectations in the early years and new specifications for mathematics, for example, would be expected to have had some impact. It is also possible that the modularisation of the examination might contribute, somewhat artificially, to performance increments. Improved performance at earlier levels of school may be undermined by the difficulty in finding qualified mathematics teachers for A level. It also needs to be noted that the quality of preparation of students choosing to take particular university courses that require mathematics may not be a good indication of the overall quality of mathematics students at A level. Shifts in the relative attractiveness of fields of study at university can alter the nature of the candidates presenting. Those best prepared in mathematics might, for example, be more inclined to choose an economics course at present than to choose engineering as in the past. Mathematics is, therefore, one area that QCA should carefully study in an attempt to understand and interpret the nature and meaning of candidate performance distributions.

Task 4: Consider any other relevant evidence or opinion, including that relating to equivalent arrangements in other countries

A central question, then, is whether the A level examination process is as good as it can be, when taking into account the knowledge of testing and large-scale examination practices elsewhere. While no other national setting will have the same traditions or expectations as those of England, is there anything that can be learned, experimented with, and perhaps adopted from the practices of other countries? Let us consider three major categories for inquiry:

- specific options for university admissions;
- standard practices to improve accuracy;
- validity evidence.

Options for university admissions

Many other countries use similar examination procedures to determine admissions to higher education institutions. Many systems also include a component of school-based assessment, equivalent to the English 'coursework' component, in determining students' results. In most cases, comparability of these assessments across schools is sought by scaling the school assessments against the external examination results. In Victoria, Australia, all students are required to take a General Achievement Test and each school's assessments of students in each subject are compared with the results of those students in the general test. If the discrepancy between the two sets of results exceeds a threshold, external examiners visit the school to provide independent assessment of the work marked within the school.

In some systems, there are no external examinations of the type conducted for A levels. In Germany, the examinations for the *Abitur* are conducted and marked within the school although, increasingly, the examinations are set externally to the schools. Checks on a school's marking of student scripts are undertaken by external assessors, typically drawn from other schools.

In the United States, schools operate much more independently in designing their curriculums and in conducting their assessments. Higher education institutions use these school assessments to determine admissions in combination with commercially prepared admissions tests. The dominant tests are the SAT, offered by the College Board and developed by the Educational Testing Service, and the examinations of the American College Testing Program, now referred to as ACT. Since states and school districts retain control over the curriculum, these tests are designed to be essentially independent of specific curriculums. The SAT I, the most popular test, is composed of test items intended to gauge general quantitative and verbal ability rather than the achievement of specific subject matter competence related to an academic curriculum. The SAT II is offered in particular subjects, and the ACT is subject focused but, again, they are based on a general expectation of competency rather than mapping to a specific, instructed syllabus. The tests have the virtue of being relatively inexpensive and quick to score, as they are principally comprised of multiple-choice items. The research evidence suggests that these tests add relatively little power to the use of secondary school grades alone for university admissions. Furthermore, there is growing interest in the United States in the use of examinations that reflect the secondary school curriculum. The impetus for this change is to help focus students' attention on the content knowledge taught in schools needed for university entry rather than spending energy to prepare for one examination. The College Board's Advanced Placement Tests are, in part, beginning to play such a role. State-wide assessments for high school graduation are another attempt to focus assessment on a curriculum specification, though they typically play less of a role in university admissions. There is, thus, growing sympathy in the United States for a system more consistent with A level attributes.

In the two Australian systems without external examinations, Queensland and the Australian Capital Territory, the schools' assessments are not simply supplemented by a general achievement measure, as in the United States. They are scaled against the general measure to render them statistically comparable across schools.

A further key difference between systems is whether university admissions are based on a profile of performances or on a summary index of performances. Most systems preserve the separate results in the subjects taken by a student. In the United States, the results on the additional measures of 'scholastic aptitude' are retained as separate measures of verbal and quantitative aptitude. Only in Australia, as far as we are aware, are the results reduced to a single dimension based on an aggregate of results in a particular number of subjects. To justify such aggregation it is necessary first to seek to express all the results on a common scale, as we have described.

Standard practices to improve accuracy

A second area for inquiry is international practices related to accuracy. Here, approaches to large-scale assessment vary broadly. In the United States, for example, where decisions involve significant consequences for individuals, the marking of scripts will typically require more than one marker, especially if the marking scheme gives latitude to make a judgement about candidates' responses. This practice is also widely adopted in Australian end-of-secondary school examinations. In the examinations for the New South Wales Higher School Certificate, for example, all essays and responses to other extended-response questions, in all subjects, are double marked. Only responses to short-answer questions are single marked. Responses to multiple-choice questions are machine marked. In about 10% of double-marked responses the marks differ sufficiently for a third marker to be called on to adjudicate. In other cases, where universal double marking is not undertaken, it is standard practice to give common papers to markers at times during the process to ascertain whether they continue to mark with the desired level of agreement. It is common for the reliability (or level of agreement) among markers to be reported. In New South Wales, individual markers do not mark whole papers. They mark only particular questions on a paper. Furthermore, procedures for managing students' papers minimise the chances of all papers from a school being marked by the same markers.

Markers are typically trained before they start marking. In many cases in the United States and Australia, marking sessions are held in supervised circumstances to control the hours of marking as well as to monitor regularly the consistency of markers' assessments with the agreed marking scheme. Recently, there has been increased interest in the use of computer-based scoring systems to serve as the 'second' marker. Most available systems require an electronic version of each script and derive their scores from a set of marked scripts. Consequently, they may not be applicable to the UK setting and schedule.

In the United States, it is expected that examination questions will be tried before their operational use in order to determine whether the types of responses they elicit fall within an expected range. In some cases, the trials are conducted with equivalent populations in other places. In other cases, the trials are conducted with the intended population by embedding items being tried out for future use in a current form of the test. In England, test materials are trialled for some testing programmes but not for public examinations. Here, the dependence of the examination on a particular curriculum makes it virtually impossible to find an appropriate population elsewhere for the trial, and embedding trial material in a current examination is rendered impossible by the necessary practice of making past examination papers available for teachers and students.

Results are often reported in terms of scores rather than summary grades, such as A, B, etc. Errors of misclassification are considerably smaller on a finer-grained scale than on such grade scales. Estimates of misclassification error are increasingly available. This statistical analysis produces an estimated likelihood that a student (or a school) placed in a classification might actually belong in another classification. For the most part, such procedures require far more time than is allocated in the English system, from sitting the examination to reporting results.

More analysis of the examination properties than is presently undertaken could be done if results at the level of individual questions were entered in the computer records for each student. That could readily be done if markers, for example, recorded their marks for each question on a machine-readable page as well as on the student script. The psychometric properties of each examination could be routinely investigated. The 'fairness' of each examination could be investigated by analysing whether the various examination questions function in the same way for different subgroups of interest or concern (eg based on gender, ethnic group, language background, region). This is not to ask whether the various subgroups achieve equivalent results, but only to ask whether it is clear that the examinations are not biased in the sense of not meaning different things to different groups.

There is also considerable effort elsewhere around the standard-setting process, but this remains an area that is largely unsatisfactory and continues to require new approaches. Much of the standard setting (that is, the definition of boundaries between grades) involves a mix of normative data (how people have performed in the past) and judgemental information (what should be expected of an individual who shows a given level of expertise). Ideally, standards could be inferred from the performance of a particular target group; that is, how well students who were very successful in university biology performed earlier in their A levels.

It is also possible to use information on the nature of performances in examinations that is typically ignored when the focus is on the normative use of results to compare individuals and to allocate grades with a primary eye on their distribution. Syllabuses specify standards of learning and performance expected of students. An examination paper based on a syllabus gives expression to those standards. Students' responses to the examination questions provide information on the levels of performance of the students in relation to the standards. There are contemporary psychometric techniques that permit the calibration of examination tasks by difficulty and the measurement of individuals by performance level on the same scale. The nature of what students at a particular performance level know and are able to do can be inferred from the nature of the tasks calibrated at that level. This kind of analysis requires using results at the level of questions and so provides another reason for entering this level of detail in the computer records for each student.

Validity evidence

A third general area of inquiry is validity evidence. From our earlier discussion, validity addresses the extent to which results can be inferred to meet the purposes of the examination. One area undergoing substantial study is the relationship of the examination questions to the domain that they are supposed to measure. In terms of the English system the problem might be phrased as the relationship of the examination question and marking scheme to the syllabus being measured. Judging content validity entails investigating whether the questions are fully representative of the intended domain; that is, whether all important content areas are equally likely to be measured. In addition, it involves making sure that the questions and marking scheme are in fact appropriate to the standards and syllabus or, in US terms, that they are 'aligned'.

In England, where there are various forms of an examination, judgements of content validity also involve investigating the comparability of the different forms. Recently, new analytic approaches have been used in other places to investigate the degree to which an examination measures what it is claimed to measure. These investigations require careful qualitative analyses of the examinations, to determine the specific content required, background knowledge, types of cognition (or intellectual skills) and task or performance demands. Based on such analyses, it is possible to compare disparate examinations on some common metrics, such as types of intellectual skills required. There have been efforts to interview examinees during or following test sections to attempt to understand the exact processes they are using when answering a question. While such work is not conducted during the 'real' examination for obvious reasons, it is a very useful approach to make sure that the test is measuring what is intended – for instance, that the examinee knows specific information in order to formulate an answer and does not succeed because of invalid approaches.

There is also an expanding body of work that attempts to make sure that the examination does not include situations that undermine its intent. An example might be a test with a written passage that takes so long to read that the examinee has no time to compose an adequate answer. Another example might be a science question that is presented in complex, or even confusing, language and so introduces a substantial reading component into what was intended as a science task. In both of these cases, the intended inferences about performance might very well be unwarranted.

Validity evidence is also commonly published to establish the legitimacy of an examination for its purposes. For example, there are numerous studies on the degree of predictive validity of the SAT admissions examination on first year university grades in the US. There are also studies of certification examinations that compare individuals known to have different degrees of expertise to determine whether the test adequately discriminates among them. Some studies look at the relationship among different measures thought to address the same general domain, for instance mathematics ability, to judge the validity of one or more of them.

Although there are many other ways in which validity has been typically investigated, there is great sympathy for looking at the broad consequences of testing programmes and for examining the degree to which multiple purposes for programmes are compatible. Thus, policy changes involving tests and raising standards may need to be examined in the light of changes in the academic preparation of secondary students, changes in the access to the curriculum afforded to various subgroups, and desires by teachers to improve their subject matter preparation.

Other countries, as England, are moving to expand higher education access at the same time as they are trying to raise standards. It is obvious that if the system is successful, more students will attempt to enter university level and they will achieve increasingly higher levels of achievement. Unless validity evidence is gathered, it will be difficult to argue convincingly that higher scores reflect improved achievement rather than lowered standards.

Task 5: Advise QCA on whether overall quality assurance arrangements match best international practice and how they might be strengthened

The panel recognises that it has an incomplete view of the A level process and has had only a limited amount of time to inquire into the system. The panel also recognises that there are careful balances to be maintained between regulator and awarding body, and between holding to experience and tradition and embarking on new approaches. We are also mindful of the policy and practical trade-offs that occur with any mandated or encouraged change. It is our considered judgement that QCA has done a commendable job in its effort to assure quality of the A level examinations, especially as QCA is a developing organisation. In addition, it must contend with a raft of notable changes: in curriculum, examination practices, consolidation of awarding bodies, policies seeking to expand upper secondary and university enrolment, and increased school accountability, among others.

Rather than provide dictums about quality assurance practices that must be undertaken, the panel recommends that QCA adopt a research-oriented stance. To that end, we recommend a set of research topics for consideration as well as a set of policies we believe are necessary to support the continued value of the examination system.

Future research

We recommend that QCA put into place a series of investigations that will illuminate future questions about whether improved test performance actually means improved achievement. These investigations may require different data-gathering by the awarding bodies. They might also mean some research into the effects of alternative practices. Further, QCA is encouraged to think about regularly collecting validity evidence on the consequences of its A level examination processes. Specifically, we recommend that QCA engage in short-term, ad hoc and long-term investigations. Without prejudicing QCA's options, it is possible that university or private consultants might independently respond to requests for research in the short-term and ad hoc realm whereas QCA, in collaboration with awarding bodies, might well conduct in-house the longer-term studies, particularly those with substantial data requirements. These studies might have external components, such as an independent design review, advice on analysis plans and commentary on the report drafts in order to assure an objective look at the phenomena. QCA should publish an annual or biennial report related to the A level examinations in which it records its plan for quality assurance, a public version of any research findings and activities accomplished, as well as specific revisions of practices planned for subsequent intervals. We will now proceed to identify a set of particular research topics that should be investigated to provide information to improve QCA and awarding body practices. We believe the topics to be essential for research. Our sketch of a research approach is only advisory and is intended to serve as a point of departure for QCA.

Recommendations for short-term research

1. Comparability of examination questions

- Investigate the feasibility of trialling sets of examination questions and marking schemes in advance of administration.

- Conduct qualitative analyses, in two subjects, of a series of examinations and resulting scripts detailing content and cognitive requirements. Judge comparability within and between subjects of the demands of the examinations and the standards of performance expected of students.

2. Psychometric properties of examination papers

- Require the awarding bodies to enter data at the level of results by question, and to undertake and report analyses of the properties of each examination.

- Investigate the use of results at question level to develop descriptions of various performance levels in examinations and attempt, over time, to extend and enrich the descriptions for each subject.

Recommendations for potential ad hoc studies

3. Quality of marking

- Ask markers to keep track of date and time spent on marking so that an analysis of impact of time and place in the sequence on assessments can be made.

- Analyse the backgrounds of markers and the marks they give.

- Explore the conduct of on-the-spot consistency checks of markers.

- Introduce limited, experimental double marking of scripts in subjects such as English to determine whether the strategy would significantly reduce errors of measurement.

4. Uniform marks vs. grades

- Investigate the extent to which universities would be willing to use uniform marks rather than grades, to minimise the problems of misclassification at the grade boundaries as well as the pressure on the process by which grade boundaries are set and the percentages of students above and below are determined.

- Investigate what information would be most useful at the high end of the performance distribution, in particular whether uniform marks would be more useful than the grade of A.

Recommendations for long-term studies

5. Validity of A level predictions

- Investigate the validity of A level results in different subjects as predictors of performance in university courses, using as criteria completion as well as level of performance at university. Investigate variations by subject, type of student, awarding body and university. Conduct studies across universities for the same subject matter, comparing A level performance, diagnostic test performance and achievement in the university course of study.

- Compare A level examinations with those of another country in the same subject, analysing the comparative demands of the task.

Policy

QCA plays an important role in maintaining accountability at all levels of the educational system. It must manage its relationships with the awarding bodies in a way that encourages their innovation and improvement while assuring fair, accurate and timely results. QCA should encourage the awarding bodies to explore new ways of meeting clear expectations about the technical quality of the examination process, from the design of examination questions through to the reporting of results. The awarding bodies have great pressures to meet immovable schedules, which may encourage them to cleave to well-honed approaches. QCA should create incentives for the awarding bodies to improve their approaches to the technical quality of the examinations.

QCA needs a strategic plan for the future. For example, it should examine the likely use of computer support in the examination process. This could include relatively well-established approaches such as computerised administration of examinations and adaptive tests. It could also include more innovative approaches, such as the use of simulations (in science, for example), computerised scoring options, computerised reporting and computer-supported test design. These options should be evaluated on the basis of their potential to improve the validity and accuracy of the testing process as well as their potential to reduce costs in the long run. QCA should also take a more active educational role with the public, in an attempt to enhance understanding of the benefits and limits of testing programmes.

Practice and policy recommendations

- QCA should manage its role in a way that supports the examination process, exhibiting both transparency and accountability in its methods.
- QCA should work to minimise unpredictability in requirements of the awarding bodies (and of schools and students). Imposition of new requirements with unreasonable timescales should be avoided.
- QCA should continue to make every effort to conduct its reviews in a timely manner on a clearly advertised schedule.
- QCA should be aggressive in communicating with policy makers about the feasibility of their expectations, in particular when it is not possible for the system to deliver what is required on the timeline envisaged.
- QCA should employ a convening function to air issues associated with standards in key areas, such as mathematics and science.
- QCA should expand its communications programme to help the public and the profession understand the benefits and limits of its testing programmes and any modifications being introduced.

Footnotes

[1] Difference among the syllabuses of awarding bodies for a particular subject is another matter. The examinations from the different bodies could all be equally valid in how they measure the relevant syllabus, while measuring somewhat different things, which reflects the differences in the syllabuses.

[2] In this approach, the percentile ranks locate students with respect to the relevant age population, not the current student cohort. Thus, if the student cohort is 40% of the notional age group, the percentile ranks would run from 99 to 60, with the number of students at each percentile rank being 1% of the age group, not 1% of the student cohort. (This is based on the assumption that the 60% of the age group not among the students would all perform worse than those taking the examinations.) If participation rates in the examinations were to change, the range of reported ranks would be adjusted to reflect the change. For example, if the cohort of students were to equal 70% of the age group, percentile ranks would be allocated from 99 to 30. (This approach is not applied to results in individual subjects but only to students' average results, developed in a manner described in footnote 5 and the associated text.) The claim that the ranks have a constant meaning over time depends crucially on three things. First is the adequacy of the assumption that growth in participation comes exclusively from students who will perform worse than those that would have been there under previous, lower participation rates. Second is existence of clear relationship between the student cohort and a well-defined population from which it is drawn. In England, there is a well-defined age group from which A level students are drawn but there is not a single student cohort since the separate awarding bodies deal with different groups of students and, in some cases, with the same students in different subjects. Third is the representation of students' performances on a single dimension (see footnote 3) as an average or aggregate of performances in single subjects and the associated (heroic) assumption that performances across the diverse fields of study available at the end of secondary education are unidimensional.

[3] As noted earlier, this strategy requires the assumption that the results from the full set of subjects are unidimensional, or, in other words, represent only one major domain rather than a number of subdomains. The unidimensionality assumption becomes more explicit when each student's rescaled marks are then added to obtain a single Tertiary Entrance Score.

[4] Australian students take five or six subjects at the end of secondary school.

[5] In Australia, the examining boards are regionally based and students take all their subjects and examinations from their State or Territory board.

[6] The shortage of markers has been substantially exacerbated by the introduction of AS courses as the first half of an A level course. This creates examinations at the end of the lower sixth form in addition to the long-standing examinations at the end of sixth form.

[7] It should be noted that each of the English awarding bodies deals with many more students than comparable bodies in Scotland, Wales and Northern Ireland as well as those in many other jurisdictions.

Appendix 1: People interviewed by the panel

Sally Francis, Principal, Farnham College
Chris Parker, Headmaster, Nottingham High School
Richard Kemp, Headmaster, Pate's Grammar School
Paul Magnall, Assistant Principal, Stoke on Trent College

Dr Nancy Lane, Department of Zoology, Cambridge University
Joshua Walker, Student, Greenhead College
Jessica Garland, Student, Greenhead College
Dr Ken Spours, Research Officer, Institute of Education
Nicholas Woodhead, Student, Leeds University
Andrew Newton, Student, Manchester University
Christopher Kitchen, Student, Manchester University
Peter Farrimond, Student, Sheffield University
Dr John Ash, Director of Admissions, University of Birmingham
Roger Clarke, Academic Registrar, University of Reading

Margaret Murray, Head of Learning and Skills Group, CBI
Mike Tomlinson, HMCI

AQA staff

Kathleen Tattersall, Director General
John Milner, Director of Examinations Administration
Eric Magee, Chair of Examiners
Elaine Huckerby, Chief Examiner

Edexcel staff

Dr Christina Townsend, Chief Executive at the time of the research
Dr Adrian Woodthorpe, Assistant Director, Assessment Design and Standards
Mary Jones, Chair of Examiners
John Adds, Chief Examiner

OCR staff

Dr Ron Mclone, Chief Executive
Simon Sharp, Director of Policy
Jean Marshall, Chair of Examiners

QCA staff

Bill Kelly, Head of Quality Audit Division
Dennis Opposs, Head of Plans, Quality Audit Division
Alan Greig, Principal Manager, Accreditation
Angus Alton, Team Leader, Comparability General Qualifications, Quality Audit Division
Penny Crouzet, Principal Officer, Quality Assurance

Between February and October 2001, an independent panel of advisers was invited by QCA to review the quality assurance systems that are designed to maintain GCE A level standards. The panel comprised Professor Eva Baker (Chair), University of California Los Angeles and Co-Director of the United States Center for Research on Standards, Evaluation and Student Testing; Dr Barry McGaw, Deputy Director for Education at the OECD; and Lord Sutherland of Houndwood, Principal and Vice-Chancellor, University of Edinburgh.

The remit of the panel was to review the overall quality assurance arrangements for GCE A level against best international practice. They reviewed the quality assurance systems used by QCA and the awarding bodies and compared these with those used in other countries.

This report describes their findings and recommendations.

For more information, contact:
Mick Walker (020 7509 5326)
ISBN 1 85838 505 9