Setting GCSE, AS and A Level Grade Standards in Summer 2014 and 2015

[First published in 2014 but also applies to summer 2015 qualifications.]

Setting and maintaining exam standards

The awarding process by which senior examiners (also known as awarders) propose what the minimum marks should be for the grade has in essence remained the same for decades. The awarders have always used a combination of qualitative and quantitative evidence such as question papers, mark schemes and completed exam papers (scripts) from the current and previous year, data on the exams such as mean marks and standard deviations, and statistical information based on the previous year's grade outcomes.

The awarders determine the minimum mark that carries forward key grade standards[1] from the previous year and is worthy of the grade. The remaining grades are set arithmetically. The assumption underlying the process has been that if the cohort taking this year's exam is similar to last year's then the results should be broadly the same.

The current process for setting grade standards is set out in the GCSE, GCE, Principal Learning and Project Code of Practice.[2] This requires that standards for key grade boundaries are set judgementally by each exam board's awarders. There are grade descriptions or performance descriptors for the standard of work expected for the award of key grades to guide the qualitative judgements, but statistical modelling based on the ability of the cohort also plays a major role. The ability to access better statistical data more rapidly has affected the approach in recent years.

[1] Grades A, C and F at GCSE and A and E at AS and A level
[2] www.ofqual.gov.uk/documents/gcse-gce-principal-learning-and-project-code-of-practice

Introducing new exams

Maintaining grade standards is most difficult when syllabuses change. Teachers and students may have fewer resources and will have to rely on specimen papers rather than past papers. There may be new topics included in the syllabus. Students are therefore likely to be less well prepared than their immediate predecessors and so perform less well.

It is also more difficult for awarders to make judgements about the quality of work that candidates have produced in response to a new style question paper. Appendix A summarises the research evidence on the accuracy of judgements of scripts that examiners are able to make when awarding GCSEs and A levels. The conventional wisdom is that the task of judging to a precise mark, at the boundary between one grade and the next, is impossible.

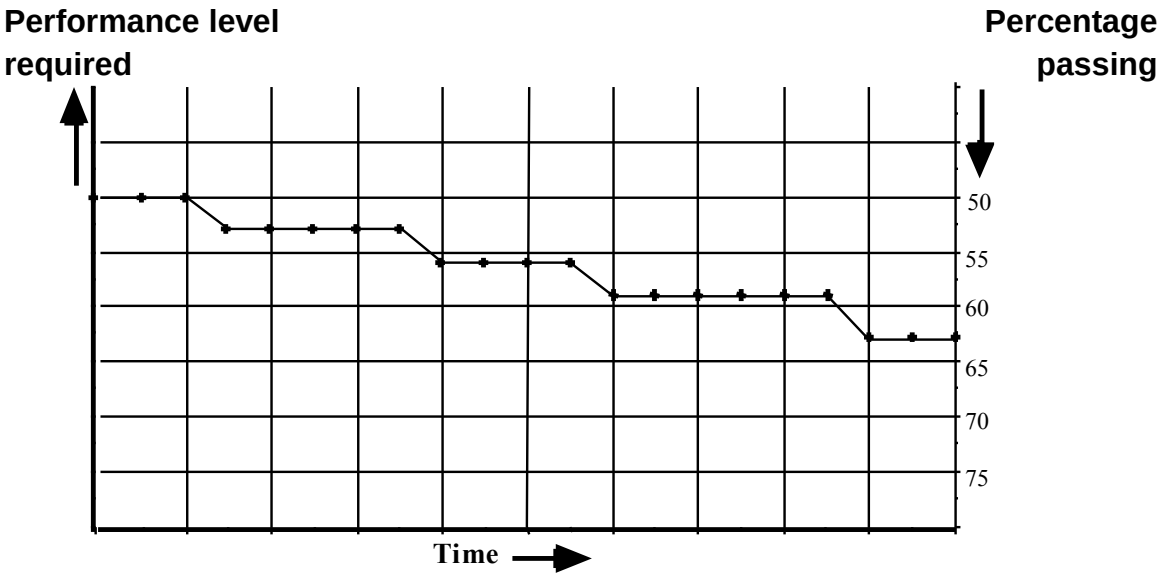
The actions that awarders take in the first year of a new exam have consequences for grade standards in the years that follow. Alastair Pollitt, a member of Ofqual's Standards Advisory Group, identified this point some 15 years ago. In discussing a change to a mathematics syllabus that occurred during 1986 he argued that in the first year the awarders had "quite properly made an allowance for the extra difficulty, accepting a lower level of performance for an A or B grade" (Pollitt, 1998).

So what happened in the following year?

> In 1987 the committee met again. This time there was no old syllabus to worry about since everyone was on the new one. This time, I suggest, they 'forgot' that a special allowance for unfamiliarity had been made last year and set the 1987 performance standard equal to the lowered 1986 one. Since then year by year comparisons have ensured that the standard set today is still set by that 'special allowance' in 1986. We might call this hypothesis 'stepwise standards' (Pollitt, 1998).

The implication of such a use of the 'special allowance' is that with successive syllabus changes the pass rate might rise but it could just be that is a function of the grade standard steadily going down. This hypothesis is illustrated in figure 1 below.

Figure 1 Falling standards (Pollitt, 1998)



New A level syllabuses

At the turn of the century there was extensive discussion between the exam boards and regulators about the most appropriate way to maintain grade standards in the first awards of the new 'Curriculum 2000' AS and A levels in 2001 – 2002. Professor Michael Cresswell, now a member of the Ofqual Board, provided much of the empirical evidence and theoretical considerations.

The regulators and exam boards decided that as a cohort, the first students should be awarded the grades that they would have received had they taken the old syllabuses so, for example, about the same proportion would be awarded a grade A. To base the awards primarily on judgements of their performance using their exam scripts would disadvantage them. This was justified on the basis of utilitarian ethical grounds as the fairest way to treat most of the candidates. This became known as the **ethical imperative** and there was an agreement to prioritise "comparable outcomes" as detailed below.

> The comparable outcomes perspective implies that grade boundaries should be fixed so as to take account of any deficits in … examination performance which are unique to the first cohort of candidates. On the other hand, the comparable performance perspective entails an acceptance that candidates' results in [the first year of a new syllabus] should suffer because for this reason they did not produce performances comparable to those which would have been achieved by candidates [in the previous year] (Cresswell, 2003).

There are good reasons to want to ensure comparable outcomes. Students who take their A levels in any particular year are competing with those from other years for access to higher education and employment. It would not be fair to one year's students if their outcomes were generally poorer simply because they were the first students to sit a new set of examinations.

This approach was also used successfully for the first awards of the revised A levels in 2010. The table below shows the proportions of students achieving grades A and E in 2008, 2009 and 2010.

A level	2008	2009	2010
Grade A (cumulative %)	25.9	26.7	27.0
Grade E (cumulative %)	97.2	97.5	97.6

When A level syllabuses have not changed

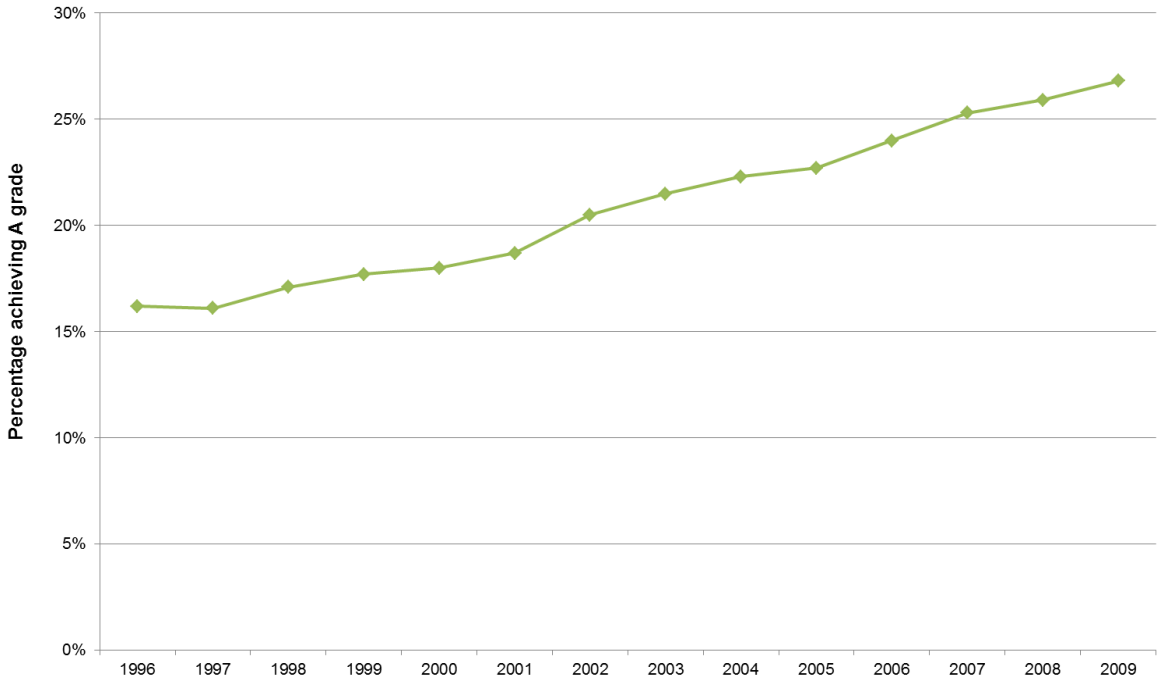
The application of the ethical imperative during the first year of a new examination then raises a fundamental question: if it is right to apply that imperative to the first year, then why should it not be applied in subsequent years?

Teaching quality and course material quality will improve gradually over a period of years. The downward adjustment of grade boundary marks during year 1 ought, in theory, to be reversed gradually during year 2, year 3, and so on, yet in practice that did not seem to happen in the early years of the last decade. Inevitably, this results in unwarranted increases in the proportions of higher grades awarded.

That suggests that there are also good reasons to prioritise comparable outcomes when the syllabuses have not changed. Following the first awards of new syllabuses where comparable outcomes are prioritised, awarding bodies had previously shifted to a varied approach, following the Code of Practice arrangements but with varying emphases on comparable outcomes or comparable performance.

We know that students' performance in examinations improves after the early years of the syllabus: teachers get used to the new requirements and there are more past papers and other resources available for students who, as a result, are better prepared and will have improved knowledge, skills and understanding (although that effect is difficult to quantify). If exam boards prioritise comparable performance over comparable outcomes, this is likely to result in 'grade drift' with, each year, gradually more students achieving each grade. Certainly A level results over time in the period before the present qualifications were introduced show a consistent rise in the proportions awarded the highest grade and this rise acts cumulatively over time.

Figure 2 A level grade A, all subjects, 1996 – 2009



DfE: students aged 16-18 at the beginning of the academic year in schools and FE sector colleges in England

The reason for this is the potential shift in emphasis from 'outcomes' to 'performance'. If an exam board selects archive scripts from the first year of a syllabus, when the focus was on producing comparable outcomes with the previous syllabus, the likelihood is that the performance on the scripts at the boundary will be at a slightly lower standard than the previous year – the last year of the established syllabus. If these archives are used to maintain standards in subsequent years, emphasising comparable performance (that is, basing decisions on judgements about students' performance), then it stands to reason that the new 'lower' grade standard will be the standard that is carried forward.

To avoid grade drift following the first awards of the new A levels in 2010, since 2011 Ofqual has required exam boards to continue to prioritise comparable outcomes as measured against the predictions based on prior GCSE achievement (see Appendix B) over comparable performance.

This is an approach that has been permitted by the Code of Practice. However, until 2011 it was not the only permitted approach. In order to make clear the emphasis on prioritising comparable outcomes to maintain as well as to set standards, the regulators strengthened the Code of Practice for 2011 to reflect this approach.

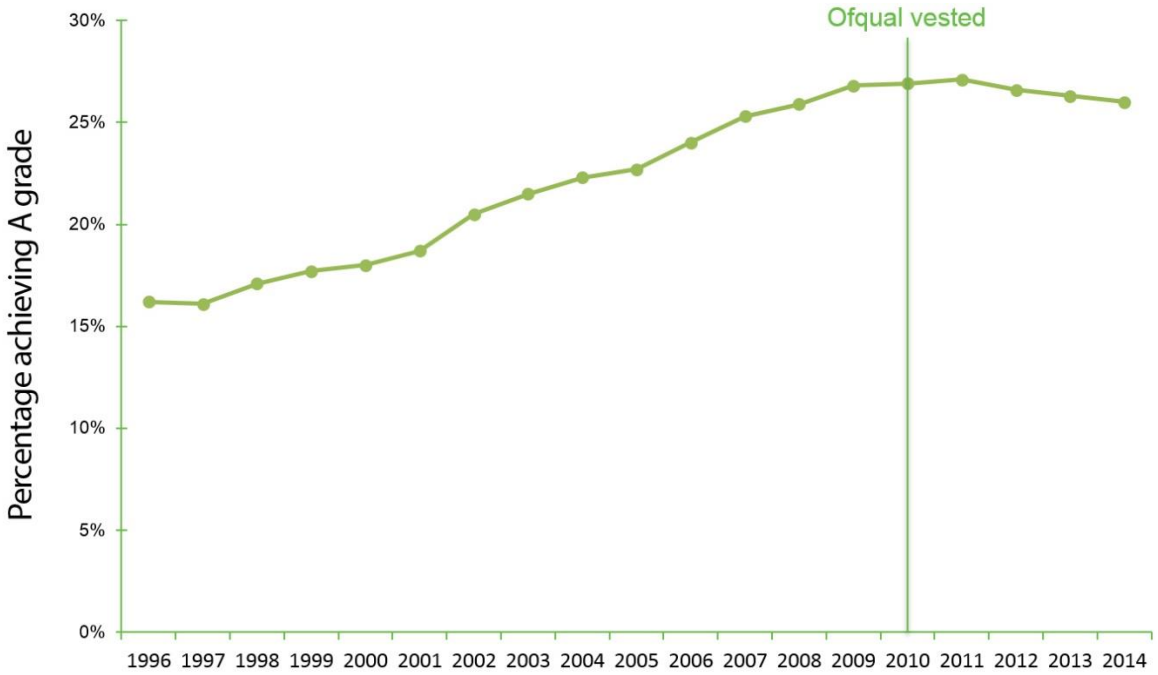
> The chair of examiners must then weigh all the available evidence – quantitative and qualitative – and recommend a single mark for the grade

> boundary, which normally will lie within the range including the two limiting marks. The choice of recommended grade boundary should be such that dependent subject-level outcomes are consistent with the evidence of relevant technical and statistical data.

In practice, this drives the final recommendations for grade boundary marks to be consistent with statistical predictions.

An updated version of the data above shows the effect that this has had since 2010. The slight dip in results in the last two years is probably due to changes in the balance of subjects that cohorts choose to study. There has been a shift recently towards what the Russell Group describes as “facilitating subjects” and what we might see as more traditional subjects.

Figure 3 A level grade A, all subjects, 1996 – 2014



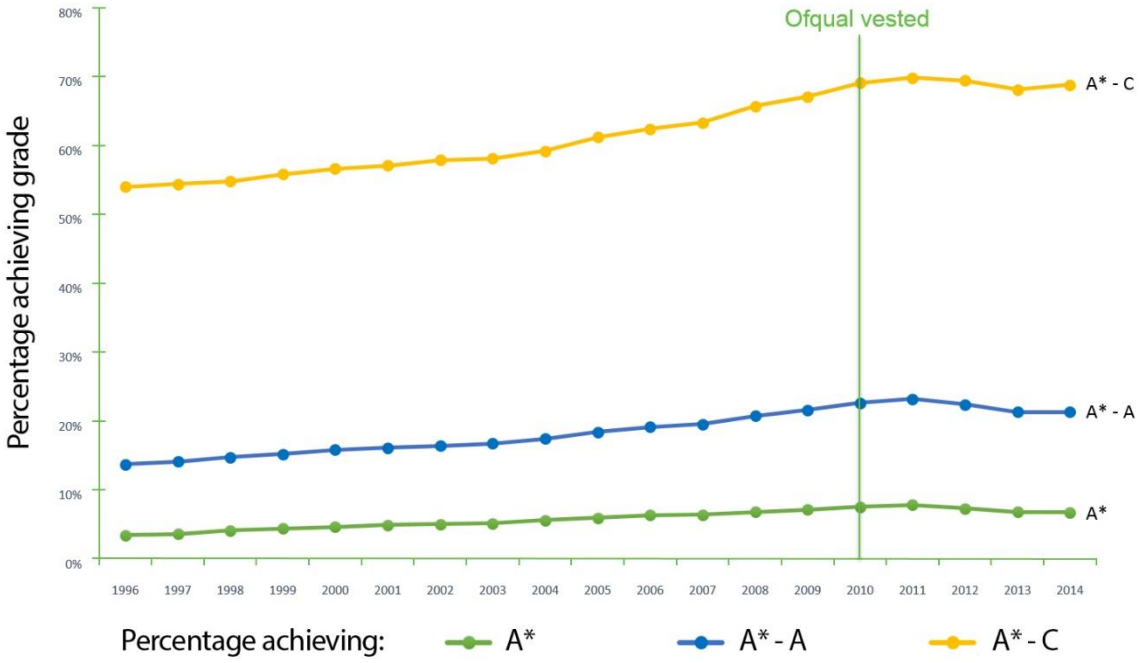
DfE: students aged 16-18 at the beginning of the academic year in schools and FE sector colleges in England

New GCSE syllabuses

In 2006 work started on a revision of GCSEs. Revised unitised syllabuses were introduced in three separate phases, the first of which involved two-year courses starting in September 2009. Before these revisions, most GCSE syllabuses were linear, in that typically all assessment was taken at the end of a two-year course.

Year on year GCSE results had shown a similar trend to that seen in A levels.

Figure 4 GCSE grades, all subjects, 1996 – 2014



The regulators and the exam boards agreed that for the new syllabuses, the exam boards should aim to produce in summer 2011 outcomes comparable with those in summer 2009. It was agreed that 2009 would be used for comparison as this was the last year in which only the previous syllabuses were available. It was also agreed that exam boards should be seeking to ensure that standards at unit level were consistent with the legacy syllabuses. In doing this they would take into account any structural changes that would impact on results (such as the impact of using a uniform mark scale)[3] but not other factors, such as any impact of students' immaturity when entering units early.

However, as the GCSE English issues in 2012 showed, implementing the focus on comparable outcomes was less straightforward at GCSE than at A level, for several reasons.

- In A level, AS denotes a clear halfway point, which provides a degree of consistency in when students take their units and it provides an opportunity for exam boards to check their progress towards outcomes that are comparable to those in previous years most unitised GCSEs had no prescribed route through the syllabus, and no 'halfway point'.

[3] In unitised qualifications units can be taken at different times. The questions papers might vary in difficulty from one sitting to another. The intention of the uniform mark scale (UMS) is to ensure that raw marks from units taken at different times receive the same value when contributing to the final grade even if the difficulty of the papers is different.

- The number of units in different syllabuses in a subject can vary at GCSE (up to a maximum of four), whereas each syllabuses in a particular subject at A level has the same number of units.
- The challenge for new GCSEs was in achieving comparable outcomes while at the same time setting consistent standards at unit level in the series leading up to the first subject awards in summer 2011. In most units, one or more awards had already been made. If standards in a unit vary between series and/or if standards between units in a syllabus vary, candidates may be advantaged or disadvantaged depending on when they take their units and/or according to where their strengths are in a subject. Schools targeting entries of particular groups of students by board and by tier adds to the challenge of making good awarding decisions.

The other question to consider is what data the exam boards use to help them succeed in achieving 'comparable outcomes'. When awarding new A levels in summer 2010 we prioritised comparable outcomes, the exam boards making adjustments to grade boundaries so that candidates were not advantaged or disadvantaged compared to their immediate predecessors because of the change in the examination structure (fewer units in most subjects) and in the task demand (the introduction of 'stretch and challenge').

Critical to the operation of the principle in the 2010 A level awards was the use of predictions based on prior achievement at GCSE for those A level candidates aged 18 years (see Appendix B). While there were also candidates of other ages they were invariably in the minority and still had to face the same changes in examination structure and task demand.

Until 2010 different exam boards made use of different statistical evidence, including data from Key Stage 3 national tests taken in England, to predict changes in the likely GCSE results for a cohort. With data from Key Stage 3 tests no longer available as the tests stopped after 2008, exam boards sought other data to use to compare the relative ability levels of the 2009 and 2011 cohorts. The replacement was Key Stage 2 test data.

In autumn 2013 we commissioned Cambridge Assessment to review the use of Key Stage 2 data to predict GCSE outcomes. We will publish the final report later in the year, but the findings suggest that the current method is fit for purpose. Predictions derived from Key Stage 2 data are highly correlated with predictions based on concurrent attainment.[4] These have been used retrospectively as a comparison,

[4] By 'concurrent attainment' we mean students' attainment in qualifications (in this case GCSEs) in other subjects taken at the same time

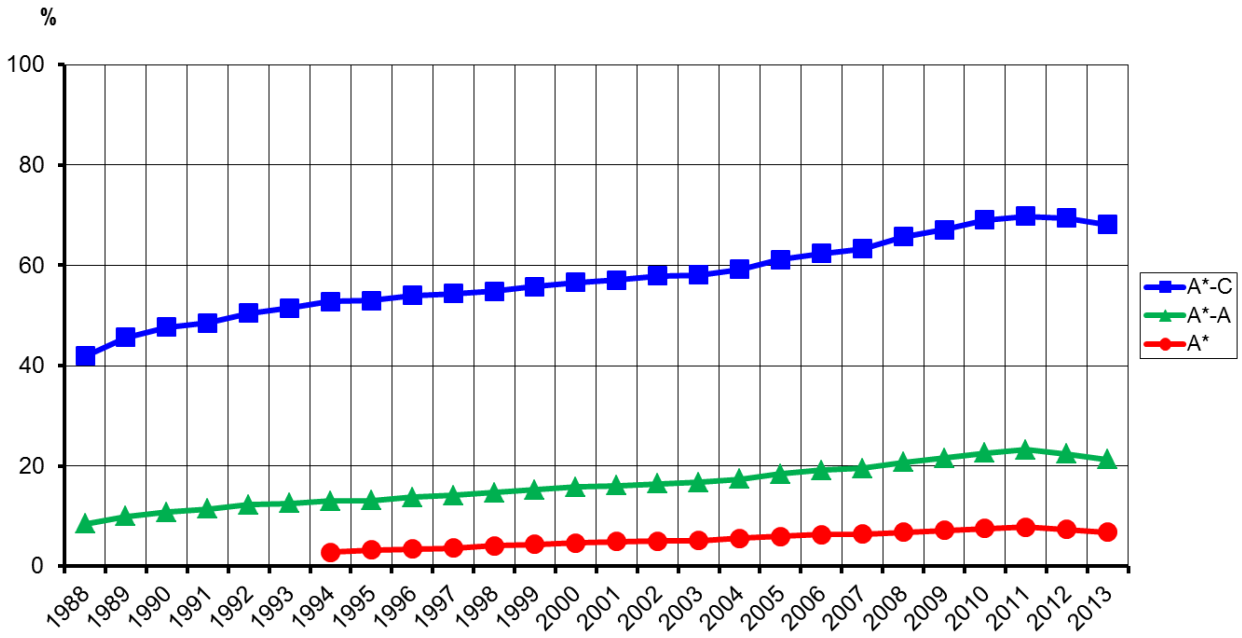
although they were not available at the time of awarding. We are discussing the detailed findings with exam boards and we will agree a common approach to the use of Key Stage 2 in generating predictions for summer 2014 GCSEs.

Having considered the issues above, we agreed with exam boards in the last three summers that emerging results in August will be reported to us against predictions in two ways:

- against predictions for the cohort based on their prior achievement at Key Stage 2, and
- as a comparison of the results achieved by common centres.[5]

An updated version of the data shown in figure 4 above shows the effect that this has had on proportions in grades. The upward trend has stopped. The fall seen in results for 2013 is largely due to the greater proportion of 15 year olds certificating. (For more detail on this, see the explanation of GCSE results[6] we published in August 2013.) These students tend to perform less well than 16 year olds.

Figure 5 GCSE grades, all subjects, 1998 - 2013



[5] A common centre is a centre that has entered students for a subject in the two years in question. The assumption is that the centre's results are unlikely to be very different in those two years, and that across the cohort as a whole, comparing results for the common centres gives an indication of whether standards between years are comparable.

[6] *A brief explanation of summer 2013 GCSE results* available at: www.ofqual.gov.uk/standards/summer-exams-2013

What we have learned from the use of comparable outcomes

While the use of comparable outcomes at A level has been little criticised of late, the same cannot be said for GCSE. This seems to have been a consequence of the different contexts within which these qualifications operate, and their different purposes.

A level results are primarily used for selection into higher education courses. The A* grade was introduced in 2010 because of complaints from a few selective universities that they were finding it increasingly difficult to sift from amongst the highest achieving candidates. From the universities' perspective, keeping the national A level grade outcomes broadly constant from year to year serves them well.

At GCSE the position is different. Schools in the state sector feel under great pressure from the Government's targets, particularly expectations that proportions of 16 year olds having achieved grade Cs in high profile subjects will rise year on year. There are currently no similar pressures for schools and colleges in relation to 18 year olds. A clear tension has arisen between Government expectations for 16 year olds' attainment and the application of the comparable outcomes approach at GCSE beyond the first year of new exams. The implication of keeping the comparable outcomes approach in years 2, 3, 4 and so on of the GCSE exams is that national grade C outcomes will remain broadly constant from year to year despite schools' increasing efforts to improve their performances.

Our position has not been that national grade C outcomes will necessarily remain the same from year to year. We say on our website:

> We believe that grade inflation – year-on-year increases in results without any real evidence of improvement in performance – should be avoided. It undermines confidence in the qualifications and in students' achievements. Our approach aims to control grade inflation, but to allow genuine improvements in performance to be recognised.

The problem lies in how the comparable outcomes approach squares with allowing "genuine improvements in performance to be recognised". This is not an issue about the first year of new exams. It is in the use of comparable outcomes in the following years.

If there is a genuine improvement in performance of students in the second year of an exam it is likely that is largely because their teachers are more familiar with the requirements of the course and the nature of the exams and so are better able to

prepare students. It is unlikely that this improved performance indicates that the latter cohort is substantially better in terms of, for example, their capacity for future learning. If we don't want unfairly to advantage the second cohort, the use of comparable outcomes appears appropriate. In doing so we should acknowledge that any small increase in, for example, students' capacity for future learning would not be recognised by increases in greater proportions of higher grades.

Suppose though that in the fourth year of a GCSE examination there are **genuine improvements in performance** of the students. Our position is that this can be an acceptable justification for the proportion of students awarded a higher grade that year to rise.

In our published process for reviewing GCSE and GCE outcome data received from exam boards we say:

> 3a Has the paper/assessment worked in a different way from previous versions? Exam boards may have evidence that the level of candidates' performance is not in line with the statistical predictions, because performance was better or worse than expected. At the award the exam scripts reviewed (at marks in the selected range for a particular grade boundary) might show that the work seen clearly merits a higher or lower grade.
>
> 3e Is there a significant mismatch between the expected and actual candidate performance? . . . We would expect convincing evidence from the exam board to support any explanation that the performance of the cohort was atypical.

The challenge arises from the nature of that evidence we require. Using only their judgement of scripts, awarders are unable to make the fine judgements necessary to decide whether a grade boundary should be put at, for example, 62, 63 or 64 marks. If that is the case, without an improbably large increase in performance from one year to the next, it is demanding for awarders' judgements to provide persuasive evidence. Indeed that is the justification for the use of a reference test to help us maintain grade standards – in a performance sense – in the new GCSEs.

As the Chief Regulator said to the Secretary of State in her 22nd August 2012 letter, our comparable outcomes approach can make it harder for genuine improvements in performance to be fully reflected in the results. It is important though that we remain open to the possibility that an exam board could present us with evidence in this

regard which, after careful consideration, we concluded did indeed justify an out of tolerance award.

Referencing grade standards

There is a misconception that some time ago A level grades were set by a system of norm-referencing – a fixed proportion of students getting each grade every year. It has never been that straightforward.

Under a norm referenced system, each student's performance is defined in relation to a norm group. It enables us to say how well a particular student has performed in relation to other students, or to the 'average' student. For example, the top 10 per cent of students receive an A, the next 30 per cent a B and so on. Depending on the assessment, the 'norm' may differ. For example, results from a particular administration of an IQ test would not indicate how well a student performed in relation to others tested at the same time, but in relation to the spread of scores that might be expected within the entire population.

Norm-referencing involves fitting a ranked list of students' raw scores to a pre-determined distribution for awarding grades. Usually, grades are spread to fit a bell curve (a normal distribution), either by qualitative, informal rough-reckoning or by statistical techniques of varying complexity.

With norm referencing, the level of performance, knowledge and skills demonstrated is not reported. The concept of a student reaching a certain acceptable 'standard' is not required.

This is in contrast to criterion referencing where a student's grade is determined by comparing his or her achievements with pre-determined performance levels or criteria. Unlike norm-referencing, a student's grade is in no way influenced by the performance of others. The challenge of criterion referenced assessments is in defining the criteria used to judge performance. In 'true' criterion referencing, the criteria must be highly specified in order to demonstrate exactly which aspects of a topic students have mastered, and only when they have mastered them all.

When assessing a very broad piece of work, however, having numerous very specific criteria is complex and unwieldy. In these cases, levels of performance may be described in holistic terms (for example in the national curriculum level descriptors). These more generic criteria require more interpretation and human judgement, which will reduce the reliability of the assessment. However, the development of criteria which would not allow a range of interpretations would be challenging and arguably these criteria would be too numerous, narrow and unmanageable.

Clearly the awarding system we use has to be valid for the types of qualifications we are considering. GCSE and A level assessments are mark-based systems so we cannot use a criterion referenced approach. It is difficult to see how we can move to a purely norm referenced approach in which there is no concept of a student reaching a certain 'standard'.

Professor Paul Newton, a member of Ofqual's Standards Advisory Group, states evidence that the system has always been based on the concept of "attainment referencing" or the "similar cohort adage". This maintains that if the nature of the cohort taking an assessment hasn't changed much year on year then the proportions of students at each grade is unlikely to change much either. He goes on to argue that this "is a rule-of-thumb that the examining boards in England have taken to heart and have integrated with their methodologies for maintaining standards". It is similar to what we now refer to as the comparable outcomes approach.

What are the alternative approaches?

We have agreed in the context of the introduction of new GCSEs that we should review the present arrangements for awarding, looking in particular at the relative contributions made by consideration of the quality of scripts and predictive statistics and how they interplay (see Appendix A, paragraphs 9 and 10). It is likely that we will find in that review that the exam boards have slightly differing views on where the balance should lie in awarding between quality of scripts and predictive statistics. Any possible changes arising from such a review – such as placing a greater emphasis on judgement of the quality of scripts – would need to be carefully evaluated and piloted before implementation, so in the short term present arrangements for awarding must continue.

In due course, at GCSE, the introduction of a national reference test will provide an additional source of evidence at awards of the new syllabuses. It may also be possible to use anchor items within exam papers to help provide evidence of performance standards over time.

As we explain above, when emerging GCSE results are sent to us each summer, the exam boards report those results against both predictions for the cohort based on their prior achievement at Key Stage 2, and as a comparison of the results achieved by common centres. So a greater emphasis on the use of the common centres' data as a prediction would be possible. However, recent analyses by the exam boards that we discussed with them in November 2013 found that predictions based on common centre data were less effective than Key Stage 2 data at predicting outcomes even when using restricted, stable centres. Further analysis is underway.

Within the present system it would of course be possible to apply the comparable outcomes approach more loosely – for example, reducing Ofqual's monitoring of live award data, widening the criteria for out of tolerance awards. That is likely to lead to a return to year on year increases in the proportions of higher grades – see figures 2 and 4 above.

Although that would satisfy some stakeholders, at GCSE there is no substantive evidence that suggests that increases in the proportions of higher grades in recent years can be justified. For example, in some respects data from the international test PISA can be treated like a reference test. In the UK the tests are taken by students in November or December of Year 11, so about six months before the end of their GCSE courses. A particular score achieved in any year should represent the same level of performance, although some statisticians have raised questions about how confident we can be about PISA data.

The data in the table below show that there has been no significant change in the UK's absolute performance in mathematics, reading or science since 2006/7.[7]

	Mean score for UK				
Year:	2000	2003	2006	2009	2012
Mathematics	Mean scores not available as the UK did not meet the sample response requirements.		495	492	494
Science			515	514	514
Reading			495	494	499

Similarly, TIMSS has been measuring trends in international science and mathematics achievement for the last 20 years on a four-year cycle. Students in England taking the test are in Year 9 so normally before they have started their GCSEs. The data in the table below show no rise in the absolute performance since 2007.

	Mean score for England				
Year:	1995	1999	2003	2007	2011
Mathematics	498	496	498	513	507
Science	533	538	544	542	533

In these circumstances from a technical perspective the present approach to guiding awarding based on the comparable outcomes approach appears the best available.

[7] In 2012 across mathematics, science and reading, there were no significant differences between Scotland, England and Northern Ireland, with the exception of mathematics where Scotland scored significantly higher than Northern Ireland. In all subjects, PISA scores for Wales were significantly below those of other UK countries and the OECD average.

GCSEs in summer 2014

We have required interim changes to English and English language (the separate reporting from 2014 of speaking and listening), to geography (awards from 2015 will be for a strengthened qualification from which 'easy routes through' have been removed) and to history and English literature (the content demand of the qualifications that will be awarded from 2015 has been increased).

We cannot know in advance how well teachers will adjust to these changes. We expect in English and English language to see increased variation in individual school performance because, in the past, schools will have behaved differently in relation to teacher marking of speaking and listening. Likewise in geography, history and English literature some schools will have exploited 'easy route' options and others will have taught the whole syllabus.

In addition, from summer 2014 the current GCSEs become linear: students must take all assessment units at the end of the course when they claim the qualification. We know that there is currently a 'route effect' in that students who have more opportunities to enter and re-enter units are likely to do better. We also know that in previous series a substantial proportion of students (approximately half in summer 2011) have taken unitised GCSEs in a linear way. Even if we could calculate an adjustment for any route effect it would be difficult to defend publicly any such adjustment, especially if it was applied to all students.

We also need to bear in mind that when we carried forward grade standards from the previous GCSEs (which were, with a few exceptions, linear) we did not make any adjustments for the facts they were unitised (other than to take account of the impact of aggregating units).

Schools will have used the GCSE qualifications in many different ways. We propose that the fairest approach to setting grade standards in 2014 and for the remaining life of these, and other current GCSEs, is to use comparable outcomes so that, if the cohort is similar, the overall proportion of each grade awarded is broadly comparable to previous years.

Modelling carried out by exam boards using data from modular GCSEs suggests that, if nothing else changes, students who have not had the opportunity to re-sit will generally do less well. However, we do not know the extent to which schools will change their approach to teaching, we cannot quantify the effects of increased maturity on exam performance and we do not know how students will respond to the changes, which will mean more teaching time. It is likely that some schools will adjust more quickly than others and that might mean some schools see greater variation in results compared to 2013 than others.

Our priority, and that of the exam boards, should be safe and fair awarding of the current qualifications at a time when we are also concerned with the introduction of new GCSEs. Our approach for 2014 and for the remaining years of the current GCSE syllabuses is therefore to continue to maintain grade standards using Key Stage 2 prediction matrices, align grade standards between exam boards and avoid grade drift.

AS and A levels in 2014

In 2014 very few, if any, changes are likely at AS and A level. We have not required any action by exam boards to make changes and, as far as we know, there are no changes planned to the accountability system that might create perverse incentives and change behaviour in schools or colleges. Higher education funding is producing an increased focus on students who can achieve grades ABB or better. We have pledged to look at some of the concerns expressed about severe grading in modern foreign languages, but we have not committed to making changes for 2014.

The only significant change is that there is no longer a January assessment opportunity for students in England (although it will still be available for students in Wales and Northern Ireland). We know that there can be a 'route effect' in unitised qualifications. However, this effect is impossible to quantify and therefore it is hard to see how we would devise and defend an adjustment. In particular, schools and colleges who had not previously entered in January would be likely to object to any blanket adjustment, even if one could be calculated.

We know from modelling work carried out by exam boards using data from AS and A levels with January awards that, if nothing else changes, students would generally do less well if only their first sitting of each unit were counted. In those scenarios, grade boundaries would have to be lowered by several marks in some cases to achieve comparable outcomes. In some subjects this could exacerbate current problems with compressed grade boundaries.

However, the modelling cannot take account of the way schools and students will adjust their teaching and learning in response to the removal of January assessment, and so the adjustments needed to grade boundaries may well be less than the modelling suggests.

Our approach for 2014 and for the remaining years of the current A level and AS qualification syllabuses is therefore to continue to maintain grade standards using GCSE prediction matrices, align grade standards between exam boards and avoid grade drift.

Appendices

Appendix A Research on the accuracy of judgements made during awarding

Appendix B Predictions for A level based on prior GCSE achievement

Appendix A: Research on the accuracy of judgements made during awarding

Cresswell (2000) analysed 108 grading decisions, comparing the boundary marks set by the examiners with those that would have been set to produce statistically equivalent outcomes. With random fluctuations in the sample of students taking examinations in any one year, it might be expected that there would be some changes in outcome and that they would reflect a normal distribution: most changes in outcomes would be small and there would be few extreme changes. Cresswell found exactly the opposite. He found few small changes: most were large swings in outcome compared with the previous year. These large swings were not explained by changes in the demographic nature of the students entered for the examinations, and they were not part of an ongoing trend.

Fortunately, the matter was not explained simply by the examiners having chosen the same boundary marks every year. There was clear evidence that examiners had responded to changes in difficulty of the examinations, with 77 per cent of the boundary marks moving in the direction predicted by the statistical evidence. In fact, examiners tended to produce boundary marks that went halfway between the previous year's boundary marks and where the statistical information suggested the boundary marks should lie.

Furthermore, there is abundant evidence that examiners are not good at discerning the difficulty of question papers. Good & Cresswell (1988) investigated examiners' ability to set grade boundaries on tests that had specifically been designed to be easy, medium and hard and which were sat by the same group of students. When students sat an easy paper, their performances were judged to be worthy of higher grades than when they sat the harder papers. The reason that there is such variability in outcomes is that examiners cannot adequately compensate in their judgements of students' work for the demands of the question papers.

Part of the awarding process has long involved reference to candidates' work on the boundary mark in the previous year. Baird (2000) investigated whether these exemplars influenced examiners' judgements in A level psychology and English by manipulating the exemplars provided to the examiners in an experiment conducted outside the operational grading process. She found that it made no difference whether examiners were given the correct exemplar for grade E or were deceived by being supplied with an exemplar for grade D. Some of the examiners were given no exemplars at all and they still set standards comparable with the other groups. Therefore, it has to be concluded that examiners are setting standards with reference not to these exemplars that they are being supplied with, but with reference to their

own mental models of the standard. There is also evidence that examiners are unduly influenced by the consistency of candidates' performances (Scharaschkin & Baird, 2000). This is an illegitimate effect because candidates are allowed to compensate for weak performances in one area with stronger performances in another in the A level and GCSE examinations. Further, examiners demonstrate a tunnel-vision effect in their judgements, as they make more severe judgements of candidates' work when they judge each question paper independently than when they judge all of their work for A level (Baird & Scharaschkin, 2002).

The awarding system also relies upon examiners being able to make qualitative distinctions between candidates' work on adjacent marks. Baird & Dhillon (2005) conducted studies with GCSE English and A level physics examiners, asking them to rank-order candidates' work in the seven-mark range in which examiners normally scrutinise candidates' work for a grade boundary decision. Care had been taken to ensure that the marking of the work included in the study was accurate. Correlations between each examiner's rank-ordering and the marks were low to moderate, and none of the 36 correlations calculated were statistically significant. None of the examiners rank ordered candidates' work well for both grade boundaries included in the study. Using a different methodology, Forster (2005) found similar results in business studies, English and geography.

This should not be interpreted as meaning that senior examiners do their job badly. On the contrary, they are selected because they are the best people for the job and show a great deal of diligence in marking and grading candidates' work in the interests of fairness. The task of judging to a precise mark, at the boundary between one grade and the next, is impossible. Candidates can reach that mark through thousands of different routes through the question paper (see Scharaschkin & Baird, 2000). Examiners are expected to be able to make a judgement about the extent to which the performances they see on the question paper are caused by a change in the question paper or in candidate preparedness. Taking these features together, there is no prototypical performance that examiners can look out for – the candidates may have reached their mark by a different, but equally valid, route or the question paper may have enhanced or detracted from their performance.

Bramley (2007) discusses the use of the paired comparison method in comparability studies. In paired comparison or rank-ordering exercises, experts are asked to place two or more objects into rank order according to some attribute. The attribute in the case of examination scripts is 'perceived difficulty'. Analysis of all the judgments creates a scale with each script represented by a number – its 'measure'. The greater the distance between two scripts on the scale, the greater the probability that the one with the higher measure would be ranked above the one with the lower measure.

Black & Bramley (2008) have argued that the rank ordering technique is a more valid use of expert judgment in awarding than the method in the Code of Practice and that it could have a role to play in providing one source of evidence for decisions on where to set the grade boundaries.

In work yet to be published, the results of earlier studies including some of those described above have been re-evaluated. It is argued that in some of those studies, data from a group of examiners shows that the group – as opposed to individuals - can distinguish between scripts with similar numbers of marks. The accuracy of these judgements of scripts by groups of examiners is then compared to the accuracy from the statistical approach that uses Key Stage 2 – GCSE prediction matrices.

These simulations suggest that in circumstances such as where there is relatively small entry, methods based on comparative judgement of scripts – particularly if the number of examiners making the judgements is far greater than is the case in traditional awarding meetings – could provide a more accurate way of setting grade boundaries than using prediction matrices as is done at present. Practical research on how the method functions in practice would be required before any such statement could be made with certainty.

The first six paragraphs of this annex are adapted from:

Baird, J, Alternative conceptions of comparability in Newton, P.E., Baird, J., Goldstein, H., Patrick, H. and Tymms, P (Eds.), (2007) Techniques for monitoring the comparability of examination standards. London: Qualifications and Curriculum Authority.

Annex B: Predictions for A level based on prior GCSE achievement

Exam boards use the relationship between GCSE performance and A level outcomes in previous years to give an indication of the overall level of achievement. This is the same methodology that is used by systems such as ALIS (2014a) and ALPS (2014b) to predict individual student outcomes based on their GCSE results. However, there is a crucial difference in that exam boards are looking at the whole cohort the entry they have for a particular subject rather than at the individual candidate level.

The exam boards work together to produce the predictions for A levels, based on the prior relationship between GCSE performance and A level performance. They then separate that out according to the entry that they have, so OCR have predictions that relate to their entry, AQA have predictions that relate to theirs, and so on.

Using these predictions means that exam boards and Ofqual can take account of differences between the entries. In any particular subject, if one exam board had an entry that comprised very high ability students, the predictions would suggest that the awarding body would have a high proportion of students achieving grade A. Expecting each exam board to have the same proportion of candidates achieving grade A might seem to be fair but it can result in some candidates being unfairly disadvantaged (or advantaged) according to the awarding body they enter with.

The predictions also provide a common measure for reporting outcomes to Ofqual in advance of results. It is important for us to be able to look across exam board, in advance of results being issued, to ensure there is a consistent approach, in the interests of fairness to candidates.

In 2009 we commissioned NFER to carry out a review of the approach. This work concluded that: “it is difficult to improve upon the current method of prediction matrices as currently applied by awarding organisations. We would recommend that this process is continued in its current form for the foreseeable future.”[8]

[8] www.ofqual.gov.uk/standards/92-articles/744

References

ALPS (2014a) - Advanced Level Performance System which provides a statistical analysis of a school or college's AS and A level results against national benchmarks. Available at: www.alps-va.co.uk (accessed 10th April 2014).

ALIS (2014b) - Advanced Level Information System which provides performance indicators for post-16 students based on GCSE data. Available at: www.cemcentre.org/alis/introduction (accessed 10th April 2014).

Baird, J., Cresswell, M.J., & Newton, P. (2000). Would the real gold standard please step forward? Research Papers in Education, 15, 213–229.

Baird, J., & Dhillon, D. (2005). Qualitative expert judgements on examination standards: Valid, but inexact. Internal report RPA 05 JB RP 077. Guildford: Assessment and Qualifications Alliance.

Baird, J., & Scharaschkin, A. (2002). Is the whole worth more than the sum of the parts? Studies of examiners' grading of individual papers and candidates' whole A level examination performances. Educational Studies, 28, 143–162.

Black, B. & Bramley, T. (2008). Investigating a judgmental rank-ordering method for maintaining standards in UK examinations. Research Papers in Education, 23(3), 357-373.

Bramley, T. (2007). Paired comparison methods. In P.E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), Techniques for monitoring the comparability of examination standards. (pp. 246-294). London: Qualifications and Curriculum Authority.

Cresswell, M.J. (2000). The role of public examinations in defining and monitoring standards. In H. Goldstein & A. Heath (Eds.), Educational standards (pp. 69-104). Oxford: Oxford University Press for The British Academy.

Cresswell, M J (2003). Heaps, prototypes and ethics: the consequences of using judgements of student performance to set examination standards in a time of change. University of London Institute of Education.

Forster, M. (2005). Can examiners successfully distinguish between scripts that vary by only a small range of marks? Unpublished internal paper, Oxford Cambridge and RSA

Good, F.J., & Cresswell, M.J. (1988). Grade awarding judgments in differentiated examinations. British Educational Research Journal, 14, 263–281.

Newton, P. E. (2011). A level pass rates and the enduring myth of norm-referencing. Research Matters, Special Issue 2, 20–26

Pollitt, A. (1998). Maintaining Standards in Changing Times. Presented at the 24th annual conference of the International Association for Educational Assessment, Barbados

Scharaschkin, A., & Baird, J. (2000). The effects of consistency of performance on A level examiners' judgements of standards. British Educational Research Journal, 26, 343–357.