CAMBRIDGE ASSESSMENT

What if the grade boundaries on all A level examinations were set at a fixed proportion of the total mark?

Tom Bramley

Paper for the Maintaining Examination Standards seminar
London, 28th March 2012

ARD Research Division

15th February 2012

University of Cambridge
Local Examinations Syndicate

Introduction

When a particular problem matters to people, and they feel that the current solution is unnecessarily complex, bureaucratic, or that it (unfairly) serves the interests of a particular group of people there is a natural tendency to ask questions of the form "Why don't we (just) …?" Some examples might be:
Why don't we just have a single … (currency / exam board / income tax rate)?
Why don't we just ban … (insert substance / activity)?

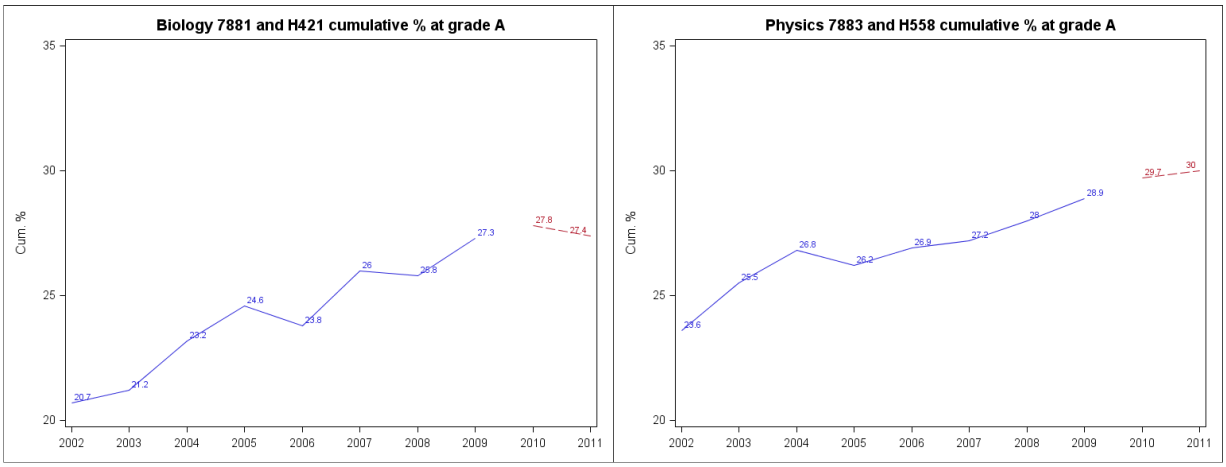
In some cases, the asking of such questions betrays naivety or ignorance, often in the form of a lack of appreciation of the likely consequences of the proposed solution. But in other cases the question, or the sentiment behind it, is not unreasonable. In these cases perhaps, those responsible for the current solution, and/or those with the power to implement a new solution, have an obligation to present a justification for the status quo (i.e. why we do it this way) and a reasoned and plausible account of what would happen if a different solution were adopted (i.e. why we don't do it that way).

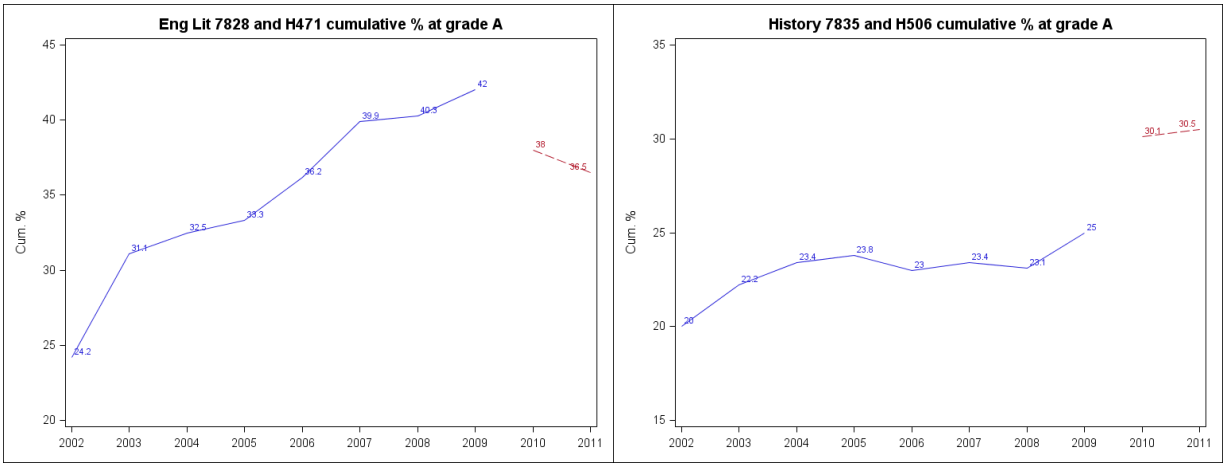
In my opinion the maintaining of examination standards is one such area. The current system is certainly complex, and the results are of great importance to a large number of people. Those of us working in educational assessment are very familiar with the kinds of headlines and stories that appear in the media relating to exam standards, and the kinds of statements put out by the awarding bodies to explain that things are more complex than they seem. I do not intend to recapitulate or summarise the issues here because they have been dealt with at great length elsewhere (see, for example, Newton et al. 2007, especially the first four chapters). Of course, there is no one single problem in the area of exam standards, but for rhetorical purposes I will imagine that the 'problem' could be expressed as "Why does it have to be so complicated?" or "Why do standards keep falling / (or rising)?".

Figure 1 illustrates the kind of data that tends to prompt the latter expression of the problem. The proportion of examinees gaining grade A in six popular mainstream academic A level subjects from OCR is shown. Discontinuities in the lines represent replacements of one syllabus (specification) with a new one.

With the exception of Psychology, there is a general trend of an increase in proportion gaining grade A over the past 10 years. The size of the increase varies according to the subject (note that the scales on the y-axes are not the same in each graph).

One simplistic solution, and the one most frequently suggested by those confronted with graphs like those in Figure 1, is "Why don't we just fix the proportion of examinees receiving each grade?" A version of this solution is proposed and defended by Paul Newton in his paper for this seminar. My purpose here is to explore a different simplistic solution to the maintenance of examination standards – namely "Why don't we just fix the grade boundaries at a particular mark?"





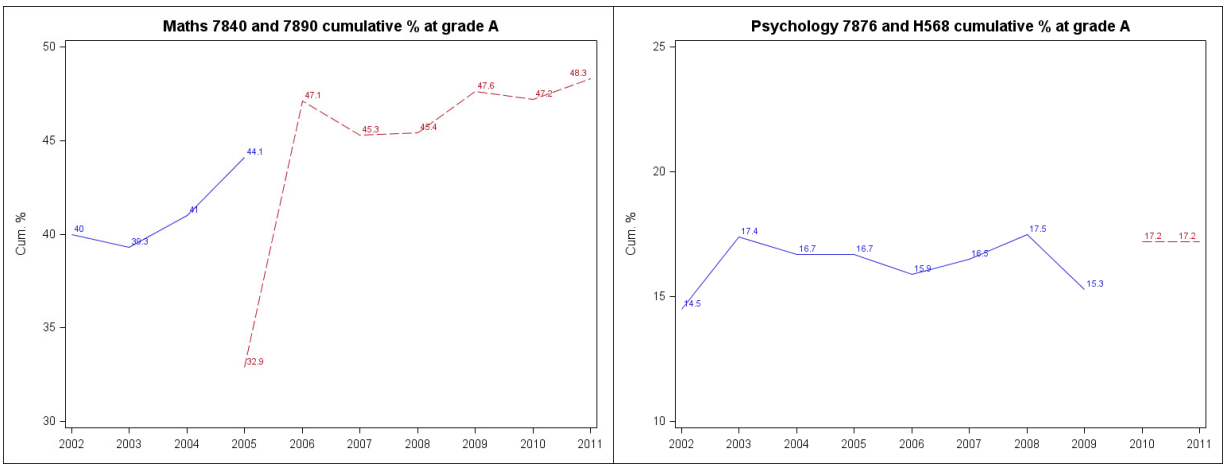


Figure 1: Percentage gaining grade A or above in six popular OCR A level specifications[1].

[1] The data underlying these graphs comes from the OCR administrative systems at the time of grading, and hence does not take into account any late arrival of marks on the system, changes to marks as a result of appeals etc. The absolute values shown (but not the trends) will therefore differ slightly from official published figures.

Conceptualisation of standards

Figure 2 below is important because I think it illustrates how a lot of people (including me) think about examination standards. However, it involves several rather slippery concepts that are explained below. Lack of agreement over the meanings of these concepts has bedevilled the debate over the years and I do not presume that everyone shares my understanding, nor do I wish to pre-empt the paper by Coe et al. for this seminar.

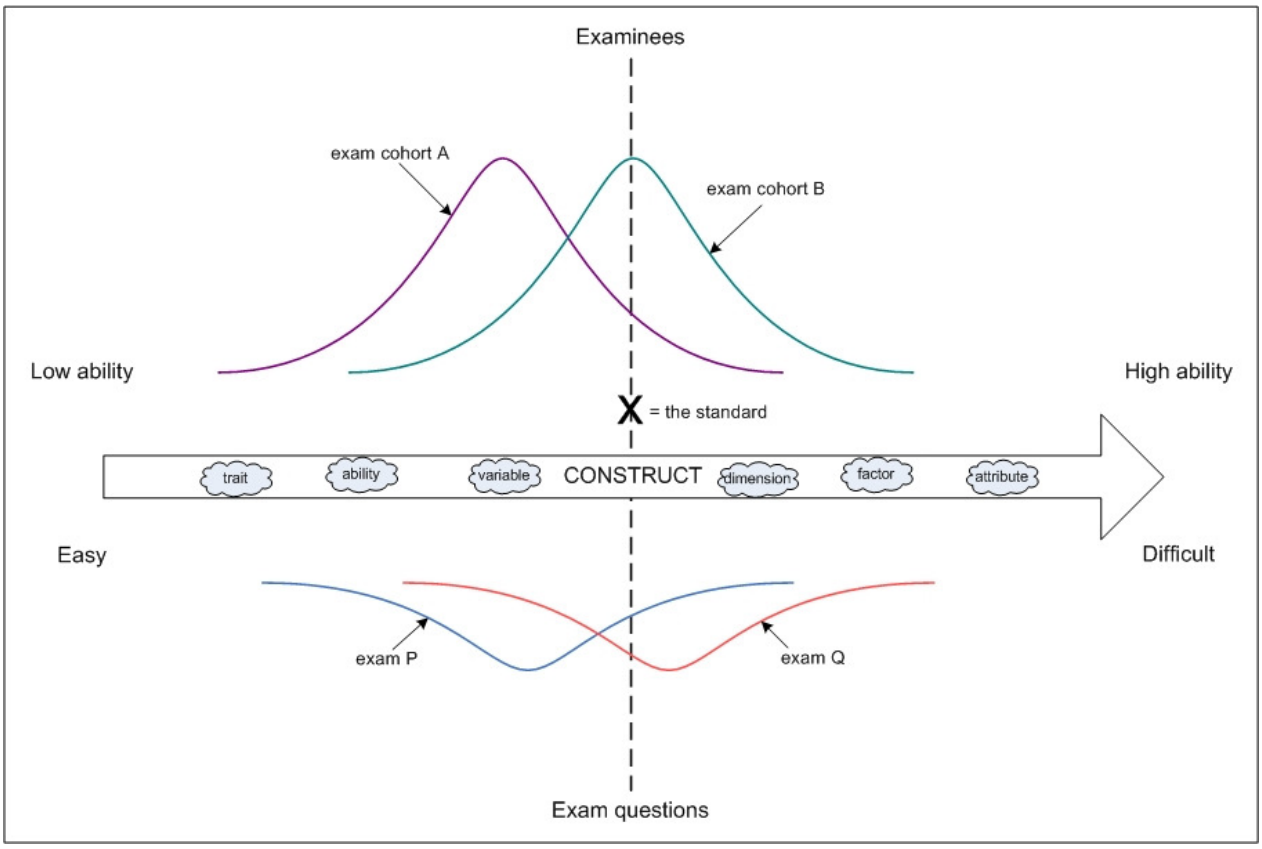


Figure 2: Schematic representation of concepts relevant to maintaining an examination standard.

The key concept, but the most difficult one, is represented by the large arrow going from left to right in Figure 2. The arrow represents whatever the assessment is supposed to be assessing. This concept has been given many names in the literature, with slightly different connotations – for example, trait, construct, ability, dimension, factor, attribute. I will use the term 'construct' here, even though I am wary of the word because it is often used ambiguously (Maraun & Peters, 2005; Borsboom et al. 2009). I choose it because it seems to fit best with how a wide range of professionals currently seem to talk about the assessment enterprise. It is conceived as an abstract line.

The second important idea is that test-takers (examinees) and examination questions (items) can be individually characterised by numbers, conceived as distances along the abstract line from an arbitrary origin (zero point). When characterising examinees, the number represents 'ability', and when characterising items it represents 'difficulty'. The term 'difficulty' is not particularly contentious, unlike 'ability', which sometimes has connotations of innate or fixed talent, or even IQ. In this paper 'ability' is simply is the word used to describe the location of an examinee on the abstract line.

The third idea is that a standard can be defined as a single point on this abstract line – 'X' marks the spot in Figure 2. To avoid the complications arising from examinations with grades (where

each grade boundary would have its own location on the abstract line), Figure 2 simply considers a single boundary – for example the A boundary on an A level examination.

These three concepts give us a simple and natural way to think about maintaining standards. The thing that does not change is the location of the standard on the abstract line – the 'X'. The things that can change are the abilities of the examinees in different cohorts[2] and the difficulties of the items in different tests. If the ability of the examinees increases (represented by the distribution of examinee ability shifting to the right along the line from cohort A to cohort B) then the proportion of them with an ability above the standard increases and the proportion obtaining grade A should increase. If the difficulty of the items increases (represented by the distribution of item difficulty shifting to the right along the line from exam P to exam Q) then the score on the examination corresponding to the standard (i.e. the grade boundary mark) should decrease.

Unfortunately, beneath this pure and simple conception of maintaining standards lurk considerable, perhaps insurmountable, problems. The construct, ability and difficulty are all defined in terms of each other. For example, to explain what it means for one cohort to have more ability in the subject than another, we might imagine the two cohorts taking the same examination, and one scoring higher on average than the other. Similarly, to explain what it means for one examination to be more difficult than another, we might imagine the same cohort taking both examinations and scoring higher on one than the other.

The joint definition of ability and difficulty finds natural expression in the mathematical equations of item response theory (IRT) models, and the large literature on test equating using these (and other) models explains how standards can be maintained within this conceptual framework. All these models rely on a link between two tests created either by data collection design (common items or common examinees), or by assumptions about ability or difficulty (see, for example, Kolen & Brennan (2004).

In A level examinations, expense and security concerns preclude the pre-testing of items, and concerns about practice on past papers prevents item re-use. Re-sit examinees are not considered to form a common link between two tests because the assumption that their ability has not changed is not tenable.

So if examinee ability can change from one cohort to another, and if item difficulty can change from one exam to another, and if the construct and standard are mere abstractions anyway – how can we know where to set the grade boundary on a given exam? This is the problem that the current complex procedures attempt to solve.

The main difficulty is that we do not have any agreement on what the criterion for a successful solution is. In other words, how do we know when standards have been maintained? In effect the procedures that are used come to *define* what it means to maintain a standard. This is not simply an undesirable feature of the current procedures, it is intrinsic to the standard maintaining problem.

The current (complex) solution

I agree with Paul Newton (2011) that what he has dubbed the 'Similar Cohort Adage' is the primary assumption (definition) behind the current procedures. That is, if there is no reason to think that the current cohort of examinees is of different ability to the previous cohort then there is no reason to expect the distribution of grades to differ.

This means that the primary orientation of current procedures is to get some kind of fix on the distribution of examinee ability – the top part of Figure 2. While once this was done using school

[2] 'Cohort' here means the group of examinees entering for the examination in a particular subject specification from a particular board in a particular session. It does not refer to the group of examinees taking all subjects across all boards (cf footnote 3).

type as a proxy for ability (on the assumption that a higher proportion of examinees from independent and selective schools implied a higher ability cohort and vice versa), nowadays this is done more directly using a measure of prior attainment. In the case of A levels this is the mean GCSE score. The assumption (definition) is now that if an examination cohort contains a higher proportion of examinees from the top deciles of the overall distribution of mean GCSE score then it is of higher ability, and vice versa. The availability of large longitudinal databases tracking the performance of all examinees over time has enabled the development of an algorithm (the 'prediction matrices' approach) that can generate an expected ('putative') distribution of grades on an examination by applying the same relationship between mean GCSE and A level grades obtained in the previous session to the distribution of mean GCSE in the current session. A more detailed explanation of this process is given in Benton & Lin (2011).

One criticism of this approach in terms of Figure 2 is that the 'ability' construct defined by adding up GCSE grades across a variety of subjects is not the same as the construct of ability in the A level subject being examined. In other words a different arrow has been substituted! This is not necessarily a fatal problem – it just means that we need to be careful when we say what it is that standards are being maintained with respect to. (For a good discussion of this see Coe, 2010). If the prediction matrices were the sole determinant of grade boundaries then we could say that A level standards are maintained in terms of the amount of general academic attainment (two years previously) implied by each grade[3]. In fact, prediction matrices are not the sole determinant of grade boundaries, although it is fair to say that they are extremely influential. The extent to which expert judgment does, could or should play a part is discussed in the papers for this seminar by Stringer & Wheadon, and Black & Johnson.

This rather convoluted explanation of what it means to maintain an A level standard may understandably not be to the taste of all stakeholders – hence the 'why don't we just…' complaint. In the solution defended by Paul Newton in this seminar the grade (or percentile) achieved simply implies relative standing within the cohort[4]. At a stroke, Newton has dispensed with the slippery concept of the 'construct', along with the more familiar but still not straightforward concepts of ability and difficulty. But Newton's simplistic approach is still ability-oriented, even though it does not need to mention ability explicitly. The approach explored below swings the pendulum in the opposite direction to consider what might happen with a difficulty-oriented approach.

Switching the focus from ability to difficulty

It is very noticeable that in the current system the concept of examination difficulty plays a very small and virtually insignificant role, in contrast to the concept of cohort ability. In terms of Figure 2 there is currently a large asymmetry – nearly all the attention is given to the top part. First of all, inferences are made about the relative ability of the current cohort. Once this has been done, then given the score distribution on a particular examination, inferences can be made about the difficulty of the examination. For example – 'this year's cohort is more able (has better GCSE results), but the mean and standard deviation of scores are the same as last year. Therefore this year's examination must have been more difficult, and we should reduce the grade boundaries to maintain the standard'.

Thus inferences about examination difficulty are only made indirectly once inferences about cohort ability have already been made. If we want to reduce the asymmetry in the current system we need a way of directly making inferences about test difficulty. On the face of it, this is highly desirable because as I have argued elsewhere (e.g. Bramley, 2010; Bramley & Dhawan, 2010) in theory the only valid reason for moving grade boundaries on an examination in order to maintain the standard is if there is evidence that the difficulty has changed. It seems at best

[3] And of course this raises the question of whether the GCSE standards are being maintained from one year to another.

[4] 'Cohort' now has a different meaning, in Newton's scheme, of the group of examinees across all boards taking examinations at the same level in the same (or a similar) subject.

somewhat unsatisfactory and at worst entirely circular to obtain this evidence indirectly from how well the examinees score on the examination. I remember someone saying, soon after I started working at what was then called UCLES, that 'if we really understood what we were assessing, we would be able to set grade boundaries before the exam was taken.' Somewhat naively perhaps, I still think this is a worthy goal.

A complex solution to the problem of evaluating difficulty directly involves understanding exactly what makes examination questions difficult, which involves understanding the psychological structures and processes involved in answering them. A lot of work, both qualitative and quantitative, has taken place and is continuing in several locations but most notably in North America, in the field known as 'item difficulty modelling' or 'cognitive diagnostic modelling' (e.g. Leighton & Gierl, 2007). One of several goals of these research programmes is to be able to generate items of known difficulty automatically (i.e. by computer) without the need for expensive pre-testing.

But for this seminar I do not want to consider a different complex solution. The counterpart to Newton's simplistic suggestion of fixing proportions achieving a grade within a specified cohort is to fix the grade boundaries on exam papers. That is, if the maximum raw mark available for the paper is 100, then fix the grade A boundary at (say) 80 marks, B at 70 marks etc, and do this for all subsequent versions of the same paper.

The parallel with Newton's proposal is that while the former effectively rules out inferences about the relative ability of examinees from different cohorts unless the 'user' is prepared to make an assumption about cohort ability, this proposal effectively rules out inferences about the relative ability of examinees from different cohorts unless the 'user' is prepared to make an assumption about exam paper difficulty (e.g. that two papers were equally difficult). Given that everyone recognises that exam papers do fluctuate in difficulty (even though we only find this out after the event) what possible advantage could there be in implementing a system that does not try to allow for these fluctuations?

The first advantage is transparency. Examinees will know when they take the paper how many marks they need to achieve in order to obtain a given grade.

A second advantage is the usefulness of the inferences that can be made about examinees from knowledge of what grade they have got. Whereas in the Newton cohort-referenced scheme only inferences about relative standing are justified (without additional assumptions), in this scheme any interested party can look at the question paper (and its mark scheme) and judge for themselves what a person scoring 80% (or 50% etc.) knows and can do. In other words it lends itself to criterion-referenced interpretations, not in terms of grade descriptors but as exemplified by the exam questions and mark scheme. This in turn could lead to a healthy focus on the content and skills tested, instead of the current (unhealthy?) focus on pass rates and grade distributions.

A third advantage is perceived fairness for the examinees (with the proviso that the papers are not *perceived* to differ drastically in difficulty). They will know that their grade did not depend on the achievement of any of the other examinees in the examination, nor (in the case of A levels) on the GCSE grades of the other examinees two years previously.

A fourth advantage is that it will be possible to fix the lowest boundary at a point that still requires a meaningful proportion of the marks to be obtained. In other words, it will not be possible to gain the lowest passing grade by only obtaining a few marks. A slogan that is often heard in the context of assessment is that of 'rewarding positive achievement'. I have always understood this to mean an approach to assessment that does not primarily seek to penalise errors and that is enshrined in how the questions and mark schemes are designed (e.g. with structure in the questions and some credit for partially correct responses in the mark schemes). However, a

second way of interpreting it is that 'no matter how little you know or can do, you can still get a grade'[5]. The proposed system could help to rule out this second interpretation.

A fifth advantage is that a satisfactory 'grade bandwidth' in examination papers could be ensured. That is, the undesirable feature that sometimes arises of having grade boundaries separated by only a few marks with the consequent potential for misclassification would be avoided. (See Bramley & Dhawan, 2010 for a discussion of grade bandwidth and reliability).

A potential advantage is that there might be a positive effect on teaching. The current system (and the Newton simplistic alternative) could encourage a fatalistic approach to 'difficult' topics because teachers and students know that if a question on a 'difficult' topic appears then scores will probably be lower and hence grade boundaries will also be lower. However, if teachers and students know that they are going to have to get the marks no matter what topics are tested there is an incentive to teach the difficult topics better (so that they are no longer so difficult)!

Another potential advantage (in the A level context) is that if the boundaries were fixed at the current 'UMS targets' of 80% for an A, 70% for a B … 40% for an E there would not be any need for the UMS (Uniform Mark Scale, see AQA, 2011 and Gray & Shaw, 2009). If the raw boundaries on a unit coincide with the UMS targets then the conversion of raw marks to UMS marks is equivalent to multiplication by a constant that depends on the weighting of that unit in the overall assessment. The main purpose of the UMS is to compensate for differences in difficulty among units of an assessment for aggregation purposes. In the proposed scheme by definition there will be no difference in difficulty! One consequence is that each raw mark gained in a unit will be worth the same to each examinee no matter which session that unit was taken in.

Of course, there are many disadvantages of this simplistic solution that will immediately spring to mind, but before discussing some of those I present some empirical data about grade boundaries. While much information about A level pass rates and grade distributions over time is publicly available, and is diligently gathered and collated by the awarding bodies, there seems to have been much less interest in, and research into, grade boundary locations. In the following sections I present graphs of grade boundaries with some commentary, first looking at the general location of the A and E boundary across all A level units, then looking in more detail at the boundaries on particular units over time.

[5] I am not suggesting that any assessment professional or official body has endorsed such an interpretation.

Where have the grade boundaries been in relation to the 'targets' in the current system?

Figure 3 shows the location of the grade A and grade E boundary mark on each unit/component as a percentage of the paper total mark for GCE units/components (i.e. both AS and A2) in the four examinations sessions from January 2010 to June 2011. Unfortunately there is some duplication within each chart in the figure because OCR's administrative systems give separate codes to the same portfolio units depending on method of moderation (submitted to the electronic 'OCR repository' or submitted for postal moderation). But this is a relatively minor problem and does not obscure the main message of the charts, which show that there was quite a wide range around the 'target' grade threshold at both A and E. (The charts are presented with the units arranged from left to right in increasing order of percentage at the A boundary, which is why the E boundary appears more irregular).

The majority of units/components with the grade A boundary *above* the 80% target were portfolio units (i.e. not externally marked written examinations) in applied GCE subjects like Business, ICT, Travel & Tourism, Leisure Studies, Health & Social Care, and Art & Design.

The units/components with the lowest boundaries as a percentage of maximum mark tended to be units with externally marked written examinations in Design & Technology, ICT, Physics, Business Studies and PE, but there were fewer consistent patterns in which units/components were involved here. A unit/component with a very low boundary in one session did not necessarily have one in other sessions.

The variability in the total number of units/components per session is partly due to the difference in availability between the January and June sessions, and partly due to the fact that in January and June 2010 some units in 'legacy' (i.e. old) specifications were being examined for the last time, alongside units in new specifications in the same subjects.

It is noticeable that very few units/components had an A-E bandwidth that was greater than the 'target' value of 40 percentage points, whereas many had a bandwidth that was less. Units/components with a high A boundary tended also to have a high E boundary. This suggests that it is difficult to construct units/components that both give the desired discrimination (spread of marks) across the score range and also the desired grade distribution on the current ability-based system for determining the grade boundary locations.

Even though question paper setters are encouraged to achieve 'target' grade boundaries on their exam papers there is little suggestion from Figure 3 of much progress towards achieving these targets when all units/components are considered together. The four charts within Figure 3 look very similar.

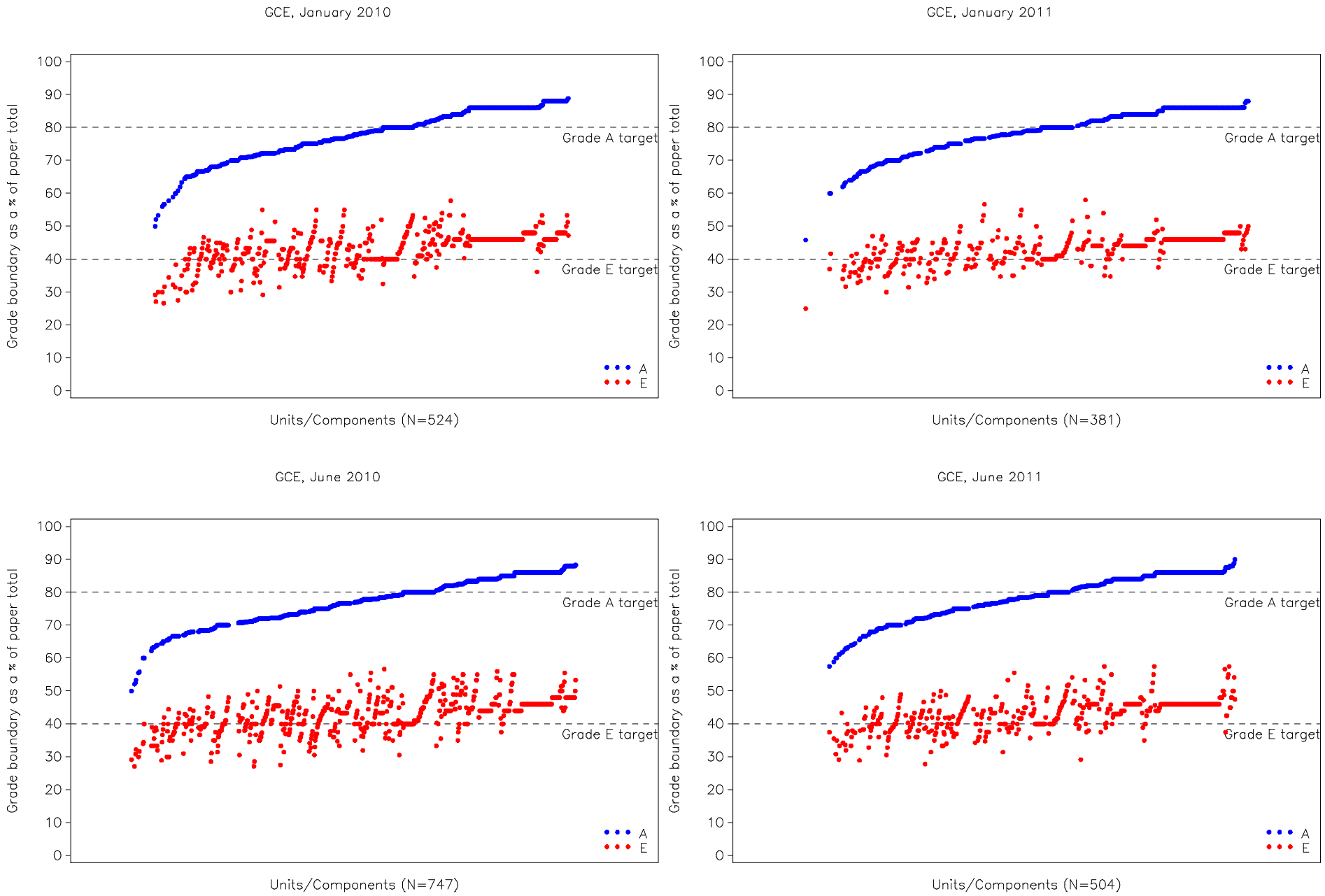


Figure 3: Grade A and grade E boundaries as a percentage of the maximum possible mark for GCE (AS and A2) units/components.

Are there any patterns in grade boundary location within a unit/component?

It is interesting to consider what (if any) patterns we might expect to see if we were to plot the location of the grade boundary within a particular unit/component of a specification over a series of examination sessions. Assuming that the papers are constructed by experts who are either intending to set papers that yield A and E boundaries of roughly 80% and 40% of maximum marks, or intending to set papers that are similar to those set previously, we might expect to see no significant patterns or trends in grade boundaries over time.

I would argue that the only legitimate patterns we should detect are a tendency for the boundary to move closer to the 'target' over time, and the occasional occurrence of a 'step-change' to reflect a deliberate decision made at a particular point in time to make an examination easier or more difficult (presumably because previous grade boundaries had been deemed to be excessively low or high, or because there was a conscious attempt to alter the standard for whatever reason).

One particularly interesting illegitimate pattern (only deemed 'illegitimate' in the absence of a satisfactory explanation) would be the presence of consistent differences between grade boundaries according to whether the examination took place in January or June. Given that examinees can enter for many GCE units in either January or June, there would seem to be no particular reason for setters to construct papers that are systematically easier or more difficult in January than June. Therefore, if such a pattern is found, one explanation is that the standard is being artificially altered in order to meet the demands of the current ability-driven standard-maintaining system. For further discussion of how features of the current system could cause disparities between January and June grade boundaries, see Black (2008).

In order to get a good sense of trends in the grade A boundary location, 'legacy' specifications in selected mainstream academic A levels were considered, because in general these ran from around 2002 to 2009, giving a large amount of data for comparison. In the graphs in Figures 4 and 5 the grade A boundary mark is plotted for each year, with January sessions shown in blue and June sessions shown in red. The grey horizontal line represents the 'target' A boundary at 80% of the maximum mark. This separation of January and June was deliberate in order to highlight visually any discrepancies between January and June boundaries.

English Literature specification 7828

Figure 4 shows grade boundaries from six of the seven units in this qualification. They give a good illustration of several of the points made above. For example units 2707, 2710 and 2712 show what we might call an 'ideal' pattern – not much variability in absolute terms, no consistent pattern in January/June boundary locations, and all boundaries close to the 80% target.

Unit 2709 was a coursework (internally assessed, externally moderated) unit. The grade boundaries on coursework units are (in theory) supposed to remain constant, but often need to be increased from time to time in order to 'control' grade distributions. Other justifications for increasing coursework boundaries are that schools are getting better at teaching/coaching their examinees to produce the right kind of work, or that they are getting better at 'playing the system' in terms of knowing how generous their marking can be before it provokes some scaling as a result of moderation. (It is worth reflecting on the implications of these justifications for what it means to maintain a standard). The plot for 2709 thus shows a typical pattern – the boundary was increased by 1 mark in June 2002 and again in June 2003. There followed several years of stability before another 1 mark increase was necessary in June 2008. As is often the case with internally assessed units, in contrast to externally examined units, the boundary was consistently higher than the target 80% mark.

Units 2708 and 2713 are interesting because they show a clear pattern of January/June differences. In both, the January boundaries tended to be consistently higher than the June boundaries, implying that for some reason the examination was consistently easier in January than June. It is particularly interesting that these units, although being out of the same raw mark total as the other units, carried more weight in the overall assessment – 20% as opposed to 15%. Thus a change to a boundary on these units would have had more effect in influencing the aggregate grade distribution, which suggests that the discrepancy between boundaries in January and June might have been an artefact of the standard maintaining procedures used rather than a genuine reflection of the relative difficulty of the examinations. If this is true, then it suggests that examinees taking one of these units in January might have been disadvantaged compared to those taking it in June.

Maths specification 7890

This maths specification was on a different development cycle from many other OCR A levels and was still current in June 2011, having run from January 2006. Figure 5 shows the grade boundaries in all of the units. Units 4721 to 4724 were compulsory 'Core Maths' units. Units 4728 and 4729 were Mechanics; 4732 and 4733 were Probability and Statistics; 4736 and 4737 were Decision Mathematics. Examinees had a variety of options in which two units they chose to supplement the four compulsory core units. They could choose any pair (4728 & 9 etc), or they could choose any two from 4728, 4732 and 4736. Ensuring comparability in some sense across these various routes would have formed a part of the standard maintaining (boundary setting) process in this specification.

Again, there are some interesting observations to be made from the pattern of boundaries in different units. There seems to be consistent January/June differences in the two compulsory (AS) units 4721 and 4722, with 4721 being in general easier (higher boundary) in June and 4722 easier in January. Units 4724 and 4733 also seem to show a January / June difference.

Although the maximum mark for these units was higher than for the English Literature units, it is noticeable how much more fluctuation there was in the grade boundaries. Perhaps this is not surprising given the nature of the two subjects – the English essay papers would have been predominantly marked according to levels-based[6] mark schemes, whereas the maths mark schemes would obviously have set out more specifically what was required in response to each question. We might conceive of 'difficulty' in English Literature A level as residing mainly in the mark scheme whereas in Maths we might conceive of it residing mainly in the questions. If there is more variety in Maths questions than in English Literature mark schemes this could explain the greater variety in grade boundaries.

However, some of the maths units showed greater fluctuations in grade boundary than others, which is puzzling. For example in unit 4724 the grade A boundary in June 2011 was 11 marks lower than the A boundary in January 2011 (and 3 marks lower than the B boundary).

Of course, some of the fluctuations could have been a result of deliberate policies to change either the difficulty of the paper or the grading standard. This was the case for unit 4736 in June 2011, where the grading standard was lowered because of evidence that it had been too high in previous sessions.

[6] Not necessarily a generic mark scheme within or across units, but more likely a banded mark scheme rewarding comparable skills and content allowing for differences in essay topic.

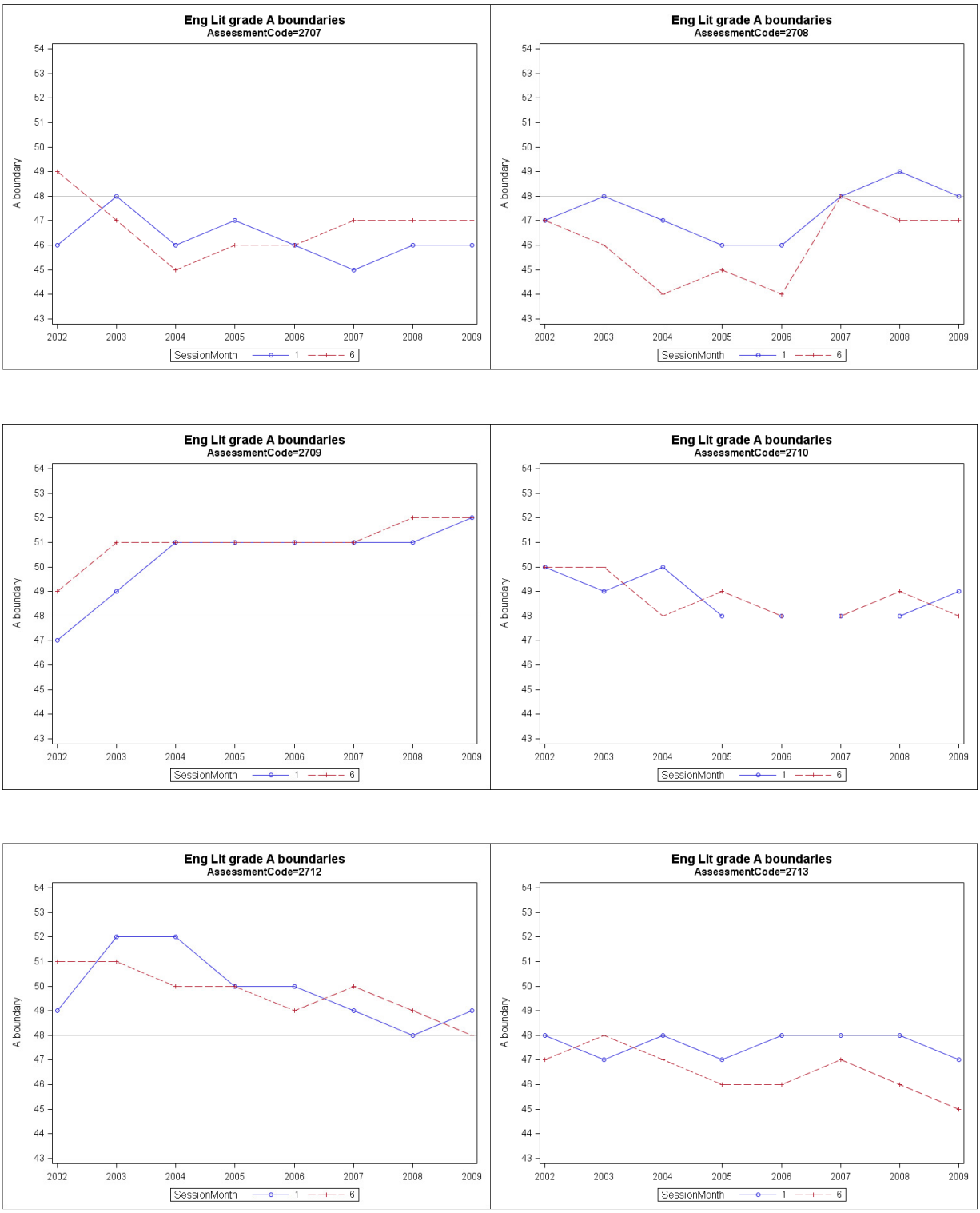


Figure 4: Grade boundaries in six of the units of English Literature (specification 7828)

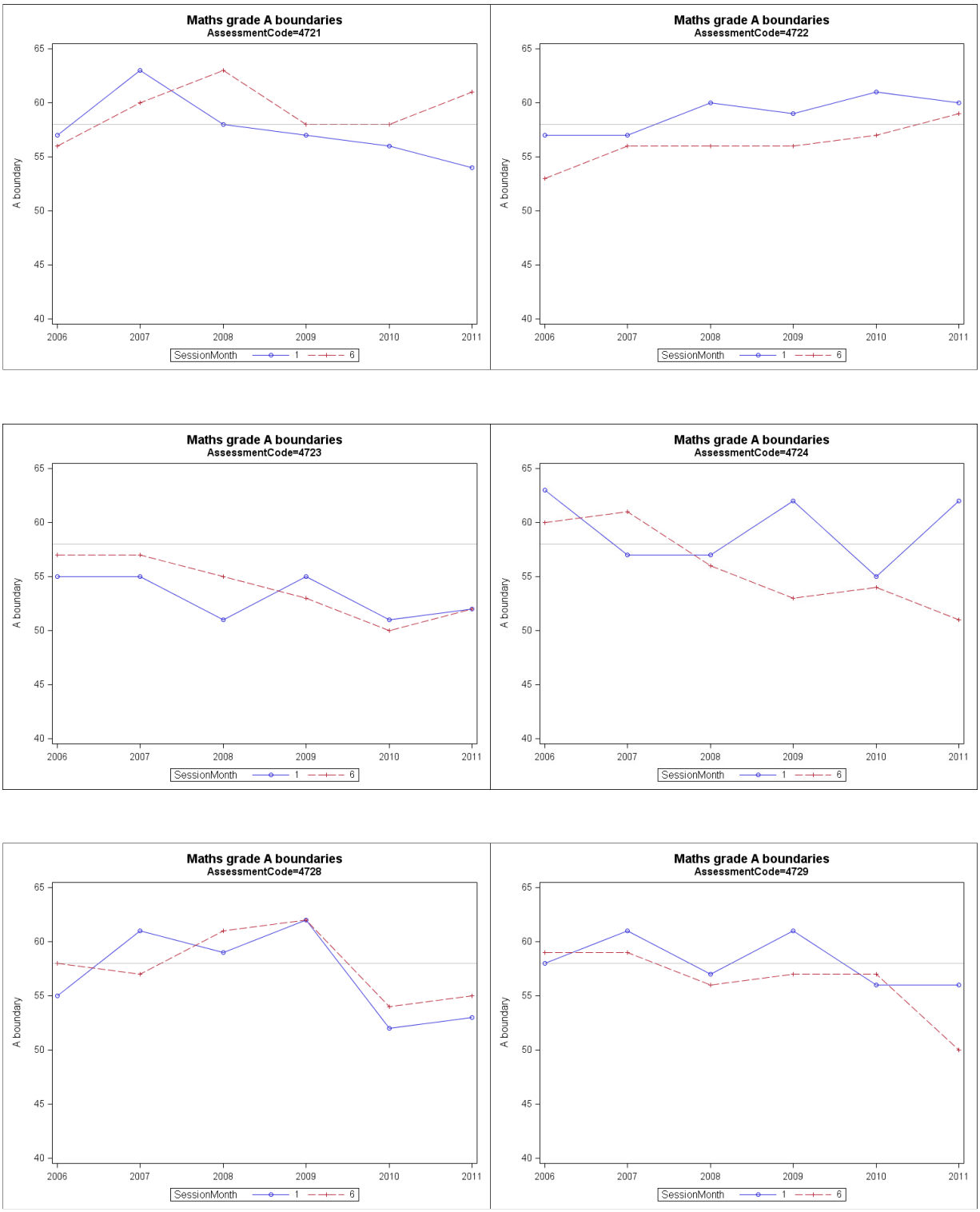


Figure 5: Grade boundaries in each unit of Mathematics specification 7828 (continued on next page).

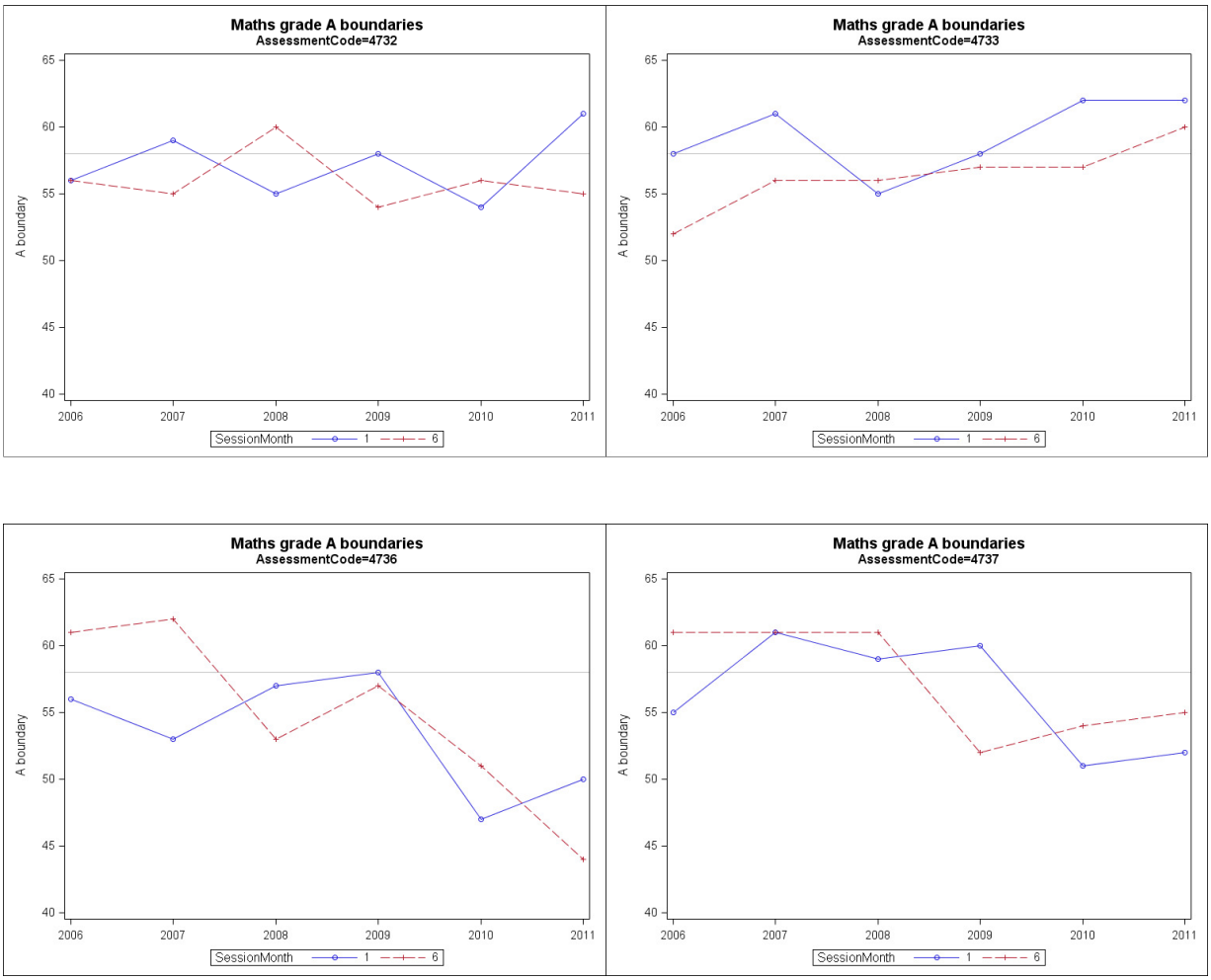


Figure 5 (contd.): Grade boundaries in each units of Mathematics specification 7828.

What might happen if the grade boundaries on all units were fixed?

To give a rough idea of what the consequences of fixing grade boundaries at their target values (80% of max mark for A, 40% of max mark for E) might be, I re-calculated the UMS scores for examinees on all units of A level maths (specification 7890), going back as far as I could using data that was readily accessible, which happened to be back to June 2007. I then calculated the new aggregate UMS scores (and hence grades) of examinees aggregating in the June 2008, June 2009, June 2010 and June 2011 examination sessions. The new (re-graded) and original UMS distributions are shown in Figure 6 below.

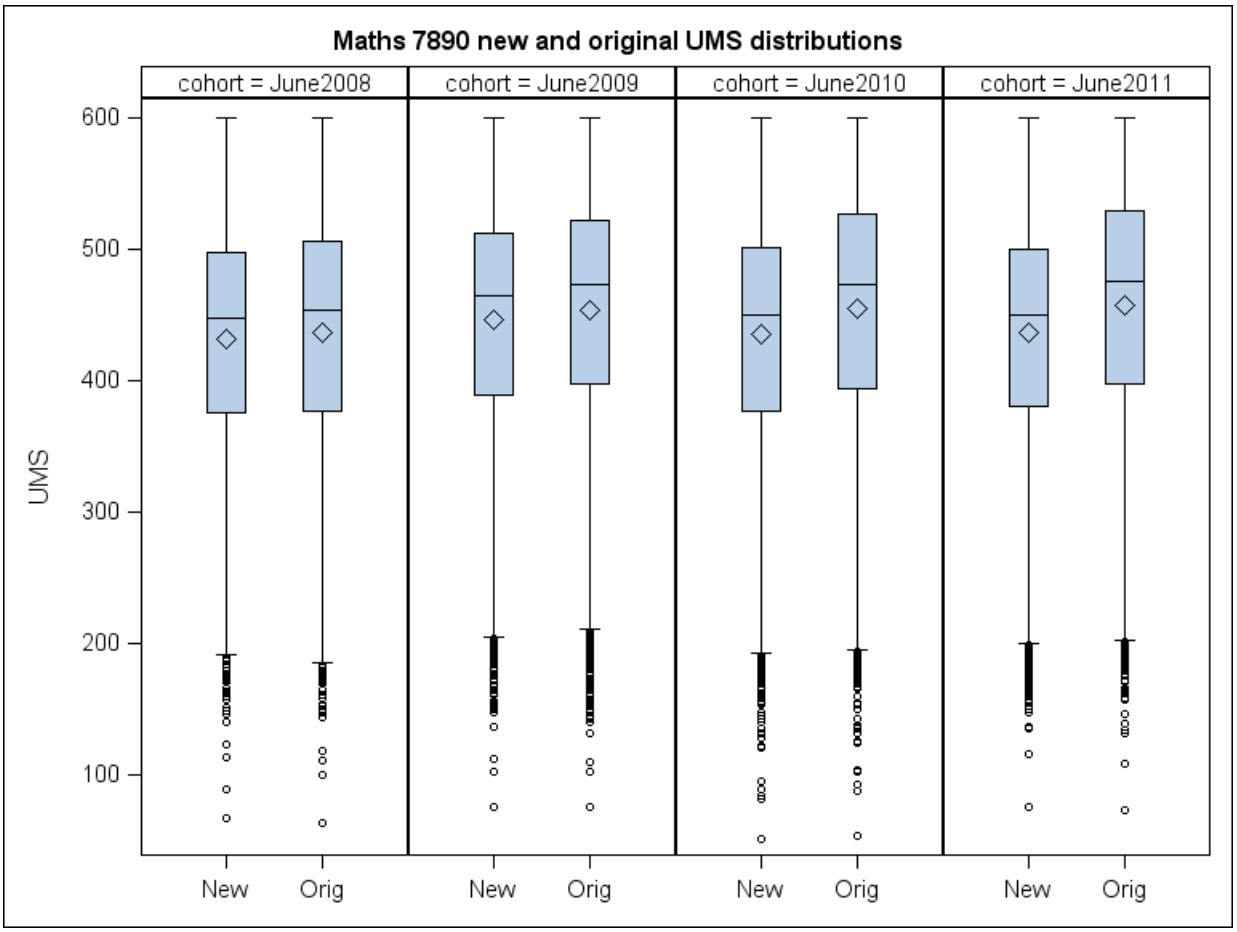


Figure 6: A level Maths (7890) original and new (re-graded) aggregate UMS distributions. The line in the middle of the box represents the median, and the diamond represents the mean.

In Figure 6 the June 2008 'original' figures are not representative because many examinees who aggregated in June 2008 had taken units in January 2007 (and the available data began in June 2007). The 2008 data is missing over 4,500 examinees as a result. The cohort sizes for the subsequent years are much more representative.

It is clear from Figure 6 that the effect of fixing the boundaries at the target values was to reduce the UMS scores of the examinees. This is not too surprising because Figure 5 showed that on most units, more boundaries were below the target 80% value than above it. The effect on the grade distributions is shown in Table 1 on the next page.

Table 1: Original and new (re-graded) cumulative grade distributions in A level maths (specification 7890).

	2008 (n=6110)		2009 (n=11216)		2010 (n=12490)		2011 (n=12972)	
Grade	Original	New	Original	New	Original	New	Original	New
A	38.2	34.7	47.4	42.8	47.3	35.4	48.5	34.9
B	61.9	60.6	68.4	65.9	68.2	61.4	69.1	61.9
C	79.1	78.9	83.7	82.5	82.9	79.4	83.9	80.5
D	90.2	90.1	93.1	92.8	92.2	90.7	93.3	92.0
E	97.0	97.1	98.1	98.0	98.0	97.4	98.0	97.6

(Note: A* grades were introduced in 2010 but were not calculated for this exercise).

If we consider just the grade distributions from 2009 to 2011 it is clear that the biggest effect of fixing the boundaries is mainly seen at the top grades (A and B). The most striking discrepancy is in the drop in percentage gaining grade A in 2010 and 2011. It is also interesting how similar the 'new' distributions are in 2010 and 2011, although it would be foolish to claim that this proposed simplistic method would be 'just as stable' on the basis of these two consecutive years (when this was clearly not the case from 2009 to 2010).

Again, the absolute values of the 'new' grade distributions should not be taken too literally[7] because the re-grading of historic units does not of course take into account what would actually have happened in terms of individual decisions to re-sit units. But it does show that the outcomes are not entirely unreasonable. If the A* had not been introduced, it might even be seen as more desirable that around 1 in 3 examinees should get the top grade rather than 1 in 2.

For interest, I repeated the re-grading process for one of the new A level specifications, Physics H558. The first units were examined in January 2008 so I was able to obtain all the data from this specification.

Table 2: Original and new (re-graded) cumulative grade distributions in A level Physics (H558)

	2010 (n=6555)		2011 (n=7732)	
Grade	Original	New	Original	New
A	29.9	8.7	30.3	9.4
B	50.4	33.8	51.9	37.3
C	69.0	60.9	70.5	65.3
D	84.4	83.5	85.9	87.4
E	95.9	97.0	96.3	98.2

Again, the biggest differences between 'original' and 'new' are at the higher grades. (Interestingly, in both years the proportion getting grade A with fixed boundaries is almost exactly the same as that getting A* in the current system).

All grades show a bigger increase in cumulative pass rate from 2010 to 2011 in the system with fixed boundaries. The problem of how to maintain standards when specifications change is an interesting one (see discussion). It is generally recognised that performance improves in the first few years of a new specification as teachers and students get accustomed to it (e.g. Pollitt, 1998). Therefore on a 'fixed boundaries' system you would probably expect to see a noticeable increase in cumulative percentages in the second year of a new specification.

[7] The re-aggregation also did not take into account what the actual selection of six units to aggregate would have been, either for the 'original' or 'new' data because in practice this selection also depends on what units are being taken for the Further Maths specification. The 'original' figures in Table were close enough (within 2 percentage points at all grades) to the actual figures to suggest that this is not a major concern for the purpose of this discussion.

Discussion

Perhaps the biggest barrier to acceptance that this simplistic proposal would have to overcome is the fact that it could potentially lead to much larger fluctuations in percentage pass rates than the current system. But this raises the interesting question of whether the pass rates fluctuate enough in the current system! Figure 7 below shows the results of random re-sampling (with replacement) ten times from an examination with an entry of 7,000 examinees and one with 1,000 examinees. It therefore gives a rough idea of the kinds of fluctuations that might be expected if the Similar Cohort Adage applied – that is, if each year's group of examinees could legitimately be considered to be a random sample[8] from the same population.

Figure 7 shows that with an entry of 7,000 the grade A percentage fluctuates approximately ±1 percentage point (around an average of about 30% – the data came from the Physics example in Table 2). With an entry of 1,000, however the fluctuations are much larger – around ± 2.5 percentage points, with a drop of 5 percentage points observed between one pair of successive points.

The current system does not really have a mechanism for dealing with random fluctuations and as a result there is the potential for 'overfitting' – producing grade distributions that match the predictions based on mean GCSE scores (or the grade distributions in previous sessions) more closely than is perhaps plausible (Bramley & Dhawan, 2011).

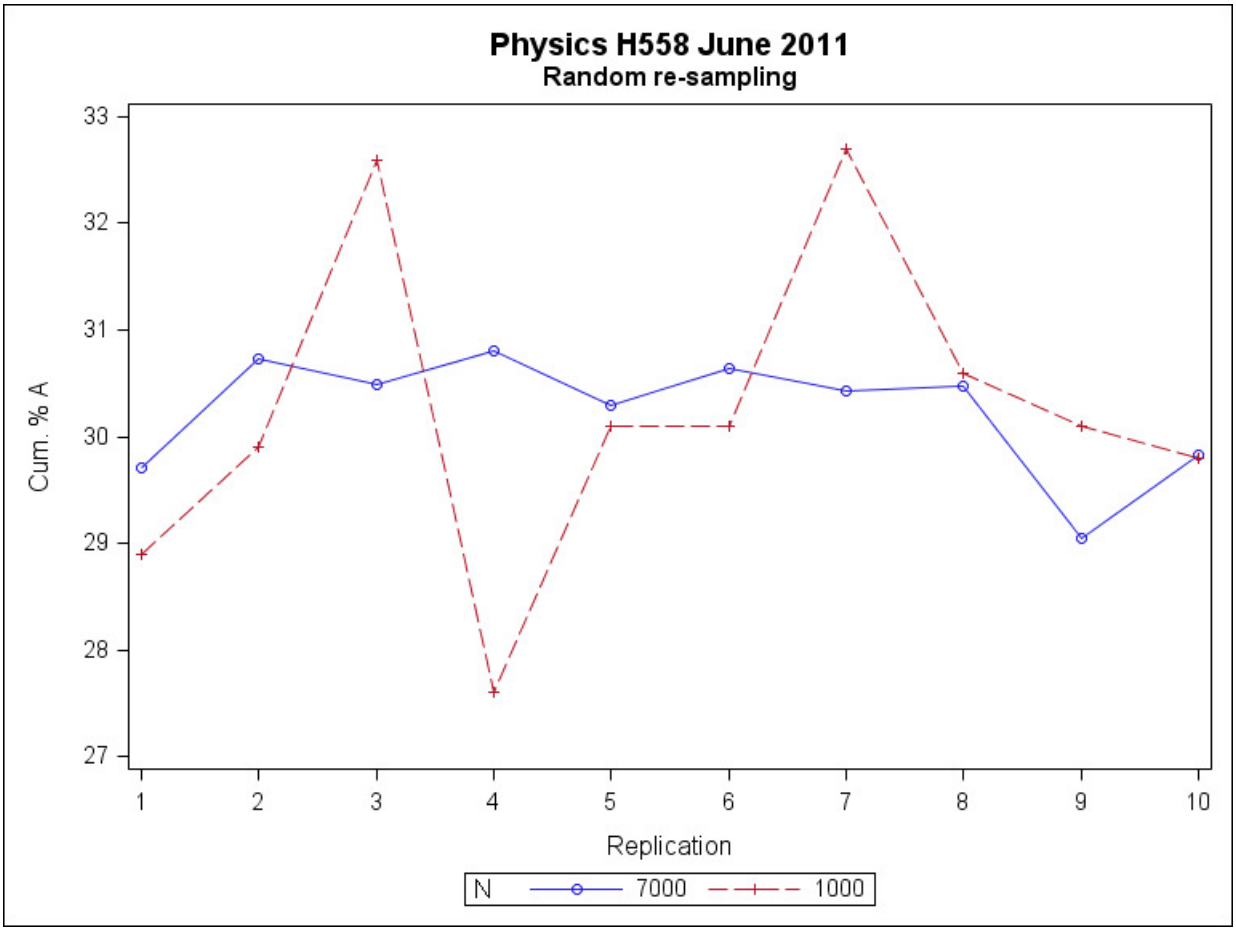


Figure 7: Hypothetical grade A pass rates for an A level with an entry of 7,000 or 1,000 examinees.

[8] A more sophisticated approach would consider both schools and examinees to be randomly sampled, which allows for the idea that examinees within a school are more likely to be similar to each other than to examinees in other schools. This 2-level re-sampling approach was also tried but the results (in this particular case) did not appear to differ substantially from those in Figure 7.

Nevertheless, would a system where the pass rates fluctuated more than they currently do be seen as unfair? Perhaps here it is worth considering a distinction between being a victim of bad luck, and being a victim of unfairness. If examination boards had explicit policies and mechanisms in place (which to some extent they already do) for attempting to ensure that successive papers were equivalent in difficulty, then they could argue that any fluctuations in difficulty could be considered to arise 'by chance'. Examinees who got a difficult paper or an easy paper might then either benefit or lose out accordingly – but this would not mean that any unfairness could be attributed. Furthermore, two factors would help to minimise the effect of luck: first the fact that most GCE A level assessments consist of 4 or 6 units. If fluctuations in difficulty were genuinely random then over a set of 4 or 6 units one would expect at least some 'cancelling out' of good or bad luck. Secondly, the opportunity to re-sit units means that people who had received a particularly difficult paper would be likely to find that their re-sit was easier.

Clearly though, an important feature of such a system would be the procedures for ensuring that, to as great an extent as possible, successive papers were of roughly equivalent difficulty. How might this be achieved? By what criterion could we agree that two papers were of equivalent difficulty? What is needed is some way of achieving the function performed by the mean GCSE scores in the current ability-focused system. Here, if two examination cohorts have the same proportions of examinees in each decile of the mean GCSE distribution, then by definition they are of the same ability. Many examination papers have an associated 'specification grid' which shows the allocation of marks to target grades and topics, and which hence provides evidence that the papers have been constructed to be of equivalent difficulty. More research effort into taxonomies of skills, and the relation of features of examination questions to their empirical difficulty would be useful, as would exploration of ways to harness expert judgment to make direct comparisons of difficulty.

Two examples of the latter have been carried out recently at Cambridge Assessment. Curcin, Black & Bramley (2010) showed that asking experts to rank-order multiple-choice questions from two different tests according to perceived difficulty (without being aware of any statistical information about facility values or score distributions) gave standard-maintaining results that were not drastically different from those produced by the official procedures. Bramley (in preparation) got groups of experts to predict where the grade boundaries would be on two GCE AS Physics units in June 2011 using only information available before the paper was taken. This information included question papers, mark schemes, specification grids, facility values from the previous June paper and mean item scores obtained by examinees exactly on the grade boundaries of the previous June paper. The eventual grade A and E boundaries (set by the official procedures) were within the range of predictions made by all expert groups (with the exception of one boundary on one unit), and for one unit the average of the predictions was equal to the eventual boundary at both grade A and grade E.

These results, while by no means suggesting that individual experts can accurately judge difficulty at the level of the individual test question, do hold out some hope that groups of experts can accurately judge relative difficulty at the level of the whole test. If ways of harnessing this expertise at the test construction stage can be found, it is not completely unreasonable to aim for a position where examination boards can demonstrate that, to the best of their ability, they have constructed papers of *roughly* equivalent difficulty *before* any examinee has actually taken the test.

Finally, I consider below some other questions that might be raised.

Would this idea (fixed boundaries) be more suitable for some subjects than others?

It seems to me that for examinations that contain mainly long essay questions marked according to levels-based mark schemes, the proposed system is actually more defensible than the current one – why should grade boundaries change if the marking criteria have remained the same? In fact, as a comparison between Figures 4 and 5 shows, there is already much less fluctuation of grade boundaries in English Literature than Mathematics. Therefore this proposal might already

be relatively painless to implement in subjects like English Literature. Furthermore, for other subjects like Art, and Design & Technology where there is a variety of optional units, in some cases the different options already have common sets of boundaries – showing that examination boards are quite capable of producing different papers with the same grade boundaries when they have to!

The greater difficulty is in constructing equally difficult papers in subjects where the 'difficulty' resides as much in the question or question topic as it does in the mark scheme – for example in subjects like maths or science. But even here there are precedents for the principle of fixed boundaries. Inspection of old A level papers from earlier decades (e.g. Elliott et al., 2008) shows that frequently they had complex rubrics with instructions like 'answer all questions in section A and any three questions from section B'. Such question choice within a paper implies an ability to construct questions of equivalent difficulty – because only one set of grade boundaries was applied to the paper. Presumably at the time, the relevant stakeholders were content to accept the good or bad luck that could arise from a particular choice of questions.

What happens when specifications (syllabuses) change?

There is often a (reasonable) presumption that in the first year or two of a new specification the level of performance is lower because teachers are unfamiliar with the new specification. In order not to 'disadvantage' examinees, often some allowance is made when grading the first examination session. In the current system, it then becomes difficult to hold back rising pass rates because it goes against the grain to penalise what appears to be genuinely better performance in subsequent sessions (cf Pollitt, ibid.). In the proposed system, this could be dealt with transparently simply by announcing that in the first session of a new specification, the fixed boundaries would be (for example) 2 percentage points lower than their intended fixed values in future sessions.

What happens when specifications (syllabuses) or assessment structures change?

This would be a problem in the proposed system – but not an insuperable one. For example, if a 6-unit specification became a 4-unit specification then we might expect to see a change in pass rates under a system with fixed grade boundaries. One can imagine a variety of 'fixes' that might be applied to smooth over irregularities when this kind of change occurs – for example a compromise model that used the current ability-focused system to adjust grade boundaries away from their intended fixed values in the first year of the new structure, but then applied the fixed boundaries in the second year, taking note in the test construction process of any units where a substantial change in the difficulty of a paper would be needed to avoid unacceptably large fluctuations in the pass rate.

How could we ensure between-subject comparability?

This is also a problem in the current system, as argued by Coe (2008). If it was desired to ensure that in some sense a grade A in Chemistry was equivalent to one in French (or, more realistically, Physics) then a definition of equivalence would need to be applied. The method suggested by Coe, or the 'subject difficulty ratings' method used in Scotland would be possible approaches.

How could we ensure between-board comparability?

This is the biggest challenge to this simplistic proposal. But it is worth noting that the current ability-focused mean GCSE method does rely either on between-board comparability existing at GCSE level, or on there being enough of a mixture in the market (schools choosing to use different boards in different subjects) at both GCSE and A level for between-board differences to cancel out.

The answer to this question that is most in the spirit of this proposal is that we could require the boards to set the same fixed boundaries, and to construct papers with the same target difficulty according to the specification grid. Then interested parties could examine the question papers

and mark schemes from each board and judge for themselves whether they consider 80% on AQA's assessment to be of equivalent worth to 80% on OCR's, for example.

Summary and conclusion

The current procedures for maintaining A level standards are complex, and lack transparency. The definition of comparability that they implement relates to an undefined construct of general academic ability that is operationalised by the mean GCSE score obtained two years previously. Although there is some indication that these procedures have in the most recent years started to check the much-commented-on yearly rises in pass rates, it is still not clear that they inspire the trust that an examination system needs.

This paper has considered a radically different approach from the current ability-driven one. It is not claimed that it would give the 'right' grade boundaries. We have seen that without a criterion for correctness this is not a meaningful aim. The content of the questions and mark scheme would exemplify what is expected in the subject, by experts and other interested parties. The papers should be constructed so that 80% for an 'A' and 40% for an 'E' are realistic and worthy goals of attainment.

By fixing grade boundaries, examinees would know before they took a paper how many marks they needed to gain to achieve each grade. They would know that their grade would not depend on the marks gained by other examinees, either in the current exam or in exams taken two years previously. A grade would have more meaning because it would more directly imply what had been achieved on the assessment. The focus of regulatory procedures and assessment research could move from sophisticated statistical modelling of the relationship between prior and subsequent attainment to a focus on content and skills tested, and to developing a better conceptualisation of what 'difficulty' means and hence to constructing papers of equal difficulty.

This proposal might even lead to desirable 'washback' effects on teaching and learning: schools would know they had to prepare their students equally well for everything that could conceivably be asked because they would know that no allowance would be made for fluctuations in difficulty.

References

AQA (2011). *Uniform marks in A-level and GCSE exams and points in the Diploma.* Version 4.2 http://store.aqa.org.uk/over/stat_pdf/UNIFORMMARKS-LEAFLET.PDF Accessed 12/12/11.

Benton, T., & Lin, Y. (2011). *Investigating the relationship between A level results and prior attainment at GCSE.* Coventry: Ofqual. Ofqual/11/5037. http://www.ofqual.gov.uk/files/2011-09-29-investigating-the-relationship-between-a-level-results-and-prior-attainment-at-gcse.pdf Accessed 12/12/11.

Black, B. (2008). *Using an adapted rank-ordering method to investigate January versus June awarding standards.* Paper presented at the fourth biennial EARLI/Northumbria assessment conference, Berlin, Germany, August 2008.

Borsboom, D., Cramer, A.O.J., Kievit, R.A., Zand Scholten, A., & Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity.* (pp. 135-170). Charlotte, NC: Information Age Publishing.

Bramley, T. (2010). *Locating objects on a latent trait using Rasch analysis of experts' judgments.* Paper presented at the conference "Probabilistic models for measurement in education, psychology, social science and health", Copenhagen, Denmark, June 2010.

Bramley, T. (in preparation). *Predicting grade boundaries using only information available before the exam has been taken.*

Bramley, T., & Dhawan, V. (2010). *Estimates of reliability of qualifications.* Coventry: Ofqual. Ofqual/11/4826. http://www.ofqual.gov.uk/files/reliability/11-03-16-Estimates-of-Reliability-of-qualifications.pdf Accessed 12/12/11.

Bramley, T., & Dhawan, V. (2011). The effect of changing component grade boundaries on the assessment outcome in GCSEs and A levels. *Research Matters: A Cambridge Assessment Publication, 12,* 13-18.

Coe, R. (2008). Comparability of GCSE examinations in different subjects: an application of the Rasch model. *Oxford Review of Education, 34(5)*, 609-636.

Coe, R. (2010). Understanding comparability of examination standards. *Research Papers in Education, 25(3)*, 271-284.

Curcin, M., Black, B. & Bramley, T. (2010). *Towards a suitable method for standard-maintaining in multiple-choice tests: capturing expert judgment of test difficulty through rank-ordering.* Paper presented at the Association for Educational Assessment-Europe annual conference, Oslo, Norway, November 2010.

Elliott, G., Curcin, M., Johnson, N., Bramley, T., Ireland, J., Gill, T. & Black, B. (2008). *Assessment instruments over time.* Paper accompanying a poster presented at the International Association for Educational Assessment annual conference, Cambridge, September 2008.

Gray, E., & Shaw, S. (2009). De-mystifying the role of the uniform mark in assessment practice: concepts, confusions and challenges. *Research Matters: A Cambridge Assessment Publication, 7*, 32-37.

Kolen, M.J., & Brennan, R.L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices.* (2nd ed.). New York: Springer.

Leighton, J.P., & Gierl, M.J. (2007). *Cognitive Diagnostic Assessment for education: theory and applications.* Cambridge: CUP.

Maraun, M.D., & Peters, J. (2005). What does it mean that an issue is conceptual in nature? *Journal of Personality Assessment, 85(2)*, 128-133.

Newton, P. (2011). A level pass rates and the enduring myth of norm-referencing. *Research Matters: A Cambridge Assessment Publication,* Special Issue 2: comparability, 20-26.

Newton, P.E., Baird, J.-A., Goldstein, H., Patrick, H., & Tymms, P. (2007). *Techniques for monitoring the comparability of examination standards.* London: Qualifications and Curriculum Authority.

Pollitt, A. (1998). *Maintaining standards in changing times.* Paper presented at the International Association for Educational Assessment annual conference, Barbados.