CAMBRIDGE ASSESSMENT

Maintaining standards in public examinations: why it is impossible to please everyone

Tom Bramley

Paper presented at the 15th biennial conference of the European Association for Research in Learning and Instruction (EARLI), Munich, Germany, 27-31 August 2013.

Research Division
Cambridge Assessment
1, Regent Street
Cambridge
CB2 1GG

Bramley.T@cambridgeassessment.org.uk

www.cambridgeassessment.org.uk

Abstract

Major reforms are currently planned for both the main high-stakes academic qualifications taken at schools in England – the GCSE (taken at age 16) and the A level (taken at age 18). The main reason for the reforms has been concern about the 'fitness for purpose' of these examinations – concern that has been fuelled by several high-profile crises. This paper shows how the increasing number of possible routes to the final grade in a subject that has arisen in the system in England over time has led to increased complexity in the mechanisms that are used to maintain standards, and the consequent undermining of confidence in the system when these mechanisms produce outcomes that are perceived to be unacceptable. The consequences of using a naïve or simplistic alternative method for setting grade boundaries (cut-scores) – namely setting them at fixed points on the raw mark scale of the assessment components – are explored in order to highlight the tensions between transparency, fairness, defensibility and logical coherence in evolving an examination system that meets the needs and expectations of its stakeholders.

Introduction

Choice is generally seen as a good thing by schools, pupils and parents. Pupils like to have some choice about which subjects they study, and teachers like to have some choice about which syllabus (within a given subject) they teach, and prepare their pupils to be examined in. The breadth of choice of subjects and syllabuses available at both GCSE and A level has increased over the past 25 years, although the number of examination boards, now known as 'Awarding Organisations' (AOs), has significantly reduced. In England, three different AOs now offer GCSEs and A levels, so the choice of a particular syllabus entails the choice of a particular AO (but in some popular subjects an AO may offer more than one syllabus). Within a particular syllabus, there may be different options and 'routes' in the scheme of assessment – for example alternative topic or content areas to be studied within science or history. Even within an individual examination, there may be some choice allowed in which questions to attempt, although the amount of choice here has over the years decreased (or disappeared entirely in some subjects). At GCSE, which is aimed at a broader range of ability than A level, in many cases there is also the choice of which 'tier' to be assessed at – 'Foundation tier' examinations are easier than 'Higher tier' examinations, but allow access to a lower range of grades: C to G and U (unclassified) as opposed to A* to E and U.

Two different assessment structures have been used at GCSE and A level – the more traditional 'linear' scheme where the various components of the assessment are taken at the end of the course, and 'modular' or 'unitised' schemes where the assessed content is broken up into discrete units which can be examined at certain points[1] throughout the course. Modular schemes were introduced into A level in the mid 1990s and widely adopted from 2002, and have more recently (but less successfully) been used at GCSE. By their nature, modular schemes introduce yet more choice into the system – schools and pupils can choose which order they teach/assess the different units, and there is the possibility for 're-sitting' units in order to boost overall grade.

The aim of the paper is to illustrate the tensions and contradictions involved in defining and maintaining standards in high-stakes academic examinations in England (GCSEs and A levels). This is done by first describing the existing complex system for setting grade boundaries (cut-scores) and illustrating some of the problems that arise. Then a radical alternative method, of great simplicity and naivety – namely using fixed grade boundaries on the raw mark scale of each assessed unit or component – is explored in order to provide an extreme contrast to the current system. The consequences of using this simplistic method in terms of year-on-year grade distributions are illustrated using data from two A level examinations (in Mathematics and Physics).

[1] There are examination sessions in January and June. However, not all units are available in the January session.

Assessment structure of a modular A level

Unit code	Unit name	Type	Max Raw	Max uniform
4721 (C1)	Core Mathematics 1	AS	72	100
4722 (C2)	Core Mathematics 2	AS	72	100
4723 (C3)	Core Mathematics 3	A2	72	100
4724 (C4)	Core Mathematics 4	A2	72	100
4728 (M1)	Mechanics 1	AS	72	100
4729 (M2)	Mechanics 2	A2	72	100
4730 (M3)	Mechanics 3	A2	72	100
4731 (M4)	Mechanics 4	A2	72	100
4732 (S1)	Probability and Statistics 1	AS	72	100
4733 (S2)	Probability and Statistics 2	A2	72	100
4734 (S3)	Probability and Statistics 3	A2	72	100
4735 (S4)	Probability and Statistics 4	A2	72	100
4736 (D1)	Decision Mathematics 1	AS	72	100
4737 (D2)	Decision Mathematics 2	A2	72	100

For a certificate candidates must have taken the following **four** mandatory units:
4721, 4722, 4723, 4724.

The other two units must be **one** of the following combinations:
4728 & 4729; 4732 & 4733; 4736 & 4737;
4728 & 4732; 4728 & 4736; 4732 & 4736.

Figure 1: The structure of one particular 6-unit A level mathematics assessment.

Students following the A level course can take the units in any session where they are available. Normally the course is completed over two years, so students 'aggregating' (see below) in June 2012 could potentially have taken units in January 2011, June 2011, January 2012 and June 2012. The AS units would normally be taken in the first year of the course (the AS qualification is 'worth' half an A level) and the A2 units in the second year of the course[2]. So, a typical pattern for the maths example of Figure 1 might be:

January 2011 – Unit 4721
June 2011 – Units 4722 and 4728
January 2012 – Unit 4723
June 2012 – Units 4724 and 4729.

But of course many other combinations are possible. Also, it is possible for examinees to re-sit units in order to improve their overall outcome. Therefore the 'typical' pattern above is actually rather untypical – Bramley & Dhawan (2012) found that only around 17% of examinees had taken the most common combination of units in one 6-unit assessment, and that 461 different combinations had been taken – more than half of them by just one examinee.

The very large number of legitimate combinations of units for examinees wishing to aggregate (i.e. 'cash in' their unit results to obtain an A level certificate) in any given examination session creates an immediate problem of comparability: how can the scores from all the different units be validly added together to give a result on the same scale? Various potential solutions to this problem were considered in the early days of modular A levels (Thomson, 1992), and the one that was adopted was the Uniform Mark Scale (UMS). A detailed explanation of how the UMS works can be found in Gray & Shaw (2009), and also in documentation on the websites of England's AOs[3]. A very brief description is given below.

[2] There is no separate qualification that just consists of A2 units.

[3] E.g. http://www.ocr.org.uk/i-want-to/do/check-results/interpreting-results/ ; http://store.aqa.org.uk/over/stat_pdf/UNIFORMMARKS-LEAFLET.PDF Accessed 13/08/13.

Description of UMS (for an A level).

The final outcome of an overall GCE A level is a grade in the following set: A*, A, B, C, D, E or U (unclassified). Each unit of a modular A level has a maximum raw mark available, and a maximum UMS mark that reflects its weighting in the overall A level. The standard-setting and maintaining procedures enshrined in the regulator's Code of Practice (Ofqual, 2011) require two cut-scores (known as 'grade boundaries') to be set on the raw mark scale of each unit. These are the grade A and grade E boundaries. The B, C and D boundaries are interpolated linearly between these boundaries[4]. The grade boundaries on the UMS are at fixed percentages of the maximum UMS available for that unit: 80% for an A, 70% for a B, … 40% for an E. So, to use the example of unit 4721 in Figure 1 (where the maximum UMS is 100), if the A boundary were set at 56 out of 72 marks, this would 'map' to 80 UMS, and a B boundary of 49 out of 72 marks would map to 70 UMS. Raw scores between the grade boundaries are mapped to the corresponding UMS scores by linear interpolation[5]. UMS scores at unit level are rounded to the nearest whole number and then aggregated. The final grade obtained depends on the aggregate UMS score. The same fixed boundaries apply, so an overall grade A is obtained by anyone with an aggregate UMS score greater than or equal to a UMS total of 80%. In the 6-unit maths example in Figure 1 this would be anyone with a score greater than or equal to 480 UMS. Likewise for grades B to E. Grade A* is an exception – this can only be obtained by examinees who have obtained a grade A overall, plus achieved an average of greater than or equal to 90% UMS on the A2 units. The A* was introduced in 2010 and was intended to increase discrimination at the top end of the scale, and to make it more difficult to achieve the highest grade by re-sitting the easier AS units that are normally taken in the first part of the course. See Acquah (2013) for further details about the A* grade.

Description of standard maintaining process

When it comes to setting the grade boundaries on the individual units, a number of different sources of evidence are taken into account by the panel of experts responsible for the process, as listed below (taken from Bramley & Dhawan, 2012):

- 'archive' scripts at the key grade boundary marks from previous sessions;
- information about the size and composition (e.g. type of school attended) of the cohort of examinees;
- teachers' forecast grades;
- the distribution of scores (mean, SD, cumulative % of examinees at each mark);
- at GCE, 'putative' grade distributions (grade distributions generated by matching examinees with their GCSE results and taking account of changes in the 'ability' of the cohort of examinees from a previous session, as indicated by changes in the distribution of mean GCSE scores;
- experts' judgments about the quality of work evident in a small sample of scripts covering a range of consecutive marks (total scores) around where the boundary under consideration is expected to be found;
- experts' judgments about the difficulty of the question paper;
- other external evidence suggesting that the particular unit/component (or assessment as a whole) had previously been severely or leniently graded and needs to be 'brought into line' – for example with other examination boards, or with other similar subjects or syllabuses within the same board.

Clearly, these different sources of evidence are more or less independent of each other, so there is the possibility for them to 'point in different directions' regarding what is the most appropriate choice for unit grade boundaries. However, in recent years one source of evidence has come to dominate the others – the 'putative grade' distributions based on the statistical relationship with prior attainment. Using this source of evidence seems to offer the best way of ensuring

[4] At a whole number of marks, following rounding rules that ensure that if unequal sizes of grade bandwidths are required, the B band is widened first, then C, then D.

[5] Raw scores outside these ranges are mapped in a similar way, with some complications (e.g. 'capping') that are not relevant to this paper. See the cited references for full details.

comparability both over time (within AOs) and between AOs. It is endorsed by the regulator (Ofqual) and indeed forms the basis on which it monitors the standard maintaining process. AOs are required to justify any departures that exceed certain tolerances from the outcomes that this method suggests. It has come to be known as 'Ofqual's Comparable Outcomes' approach in the light of recent controversies[6].

For a description of how the putative grade distributions are calculated, see Benton & Lin (2010) and Taylor (2012). Briefly – each cohort of A level examinees is split into deciles based on performance at GCSE two years previously. The distribution of A level grades achieved in a given examination in the previous year[7] is cross-tabulated against prior attainment decile – this table is known as a 'prediction matrix'. On the assumption that the grade distribution *within decile* should not change from one year to the next, the proportions achieving each grade within a decile are multiplied by the number of examinees in each decile in the current cohort to give a predicted within-decile frequency distribution, which when aggregated across the deciles gives a predicted or 'putative' grade distribution for that examination.

These 'putative' distributions apply to the A level as a whole. Each AO also produces their own putative distributions for each **unit**, but these only involve their own examinees and are not subject to regulatory tolerances. Nonetheless, they also play a prominent role in setting the boundaries on the units.

Several features of the comparable outcomes method are noteworthy:

- It requires the tracking and matching of examinees longitudinally, which requires considerable data collection effort, inter-organisational cooperation, and technological capability;
- It is a purely statistical method that does not take explicit account of either the difficulty of the examinations or the quality of examinee work;
- It produces 'targets' (which perhaps explains its desirability from the regulator's point of view).

The next section considers how the comparable outcomes approach fits into a general understanding of the logic of standard maintaining.

[6] E.g. http://www2.ofqual.gov.uk/files/2012-05-09-maintaining-standards-in-summer-2012.pdf, but note that the prediction matrices method can be applied whether or not the qualification is a new version.

[7] Or in a weighted average of previous years (Taylor, 2012)

Conceptualisation of standards
(This section is adapted from earlier work of mine – Bramley, 2012).
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 over the years bedevilled the debate about standards in England (and perhaps elsewhere!).

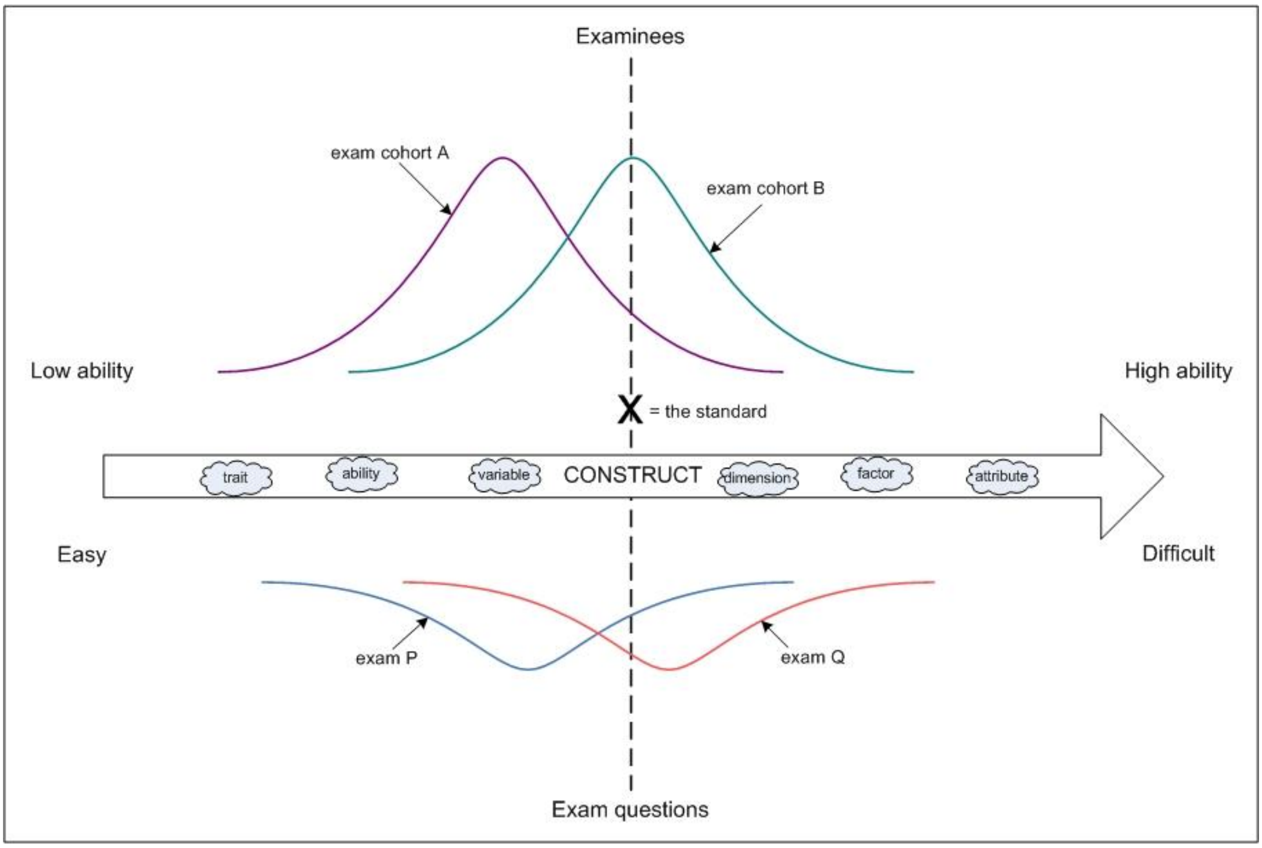

Figure 2: Schematic representation of concepts relevant to maintaining an examination standard.

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 because it seems to fit best with how a wide range of professionals currently 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 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 several grades (where each grade boundary would have its own location on the abstract line), Figure 2 simply considers a single cut-score – for example the grade 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[8] 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 primary assumption (definition) behind the current procedures is that 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. As explained above, this is done by considering the cohort's distribution in terms of deciles of prior attainment. 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.

[8] 'Cohort' here means the group of examinees entering for the examination in a particular syllabus from a particular board in a particular session.

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 lower 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, 2012) 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 somewhat unsatisfactory and at worst entirely circular to obtain this evidence indirectly from how well the examinees score on the examination.

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 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.

A simplistic 'solution' to the same problem would simply be to fix the grade boundaries on the raw score scales of each unit. 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. This would in effect ignore or dismiss, by definition, the possibility that papers differ in difficulty. Given that both experts and non-experts recognise that exam papers can and 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 did not try to allow for these fluctuations?

Advantages of fixing grade boundaries

The first advantage is transparency. Examinees would know when they took the paper how many marks they needed to achieve in order to obtain a given grade.

A second advantage is the usefulness of the inferences that could be made about examinees from knowledge of what grade they had obtained. Any interested party could look at the question paper (and its mark scheme) and judge for themselves what a person scoring 80% (or 50% etc.) knew and could do. In other words it would lend 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 would 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 would be possible to fix the lowest boundary at a point that still required a meaningful proportion of the marks to be obtained. In other words, it would 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'[9]. A 'fixed boundaries' approach 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 on the raw scale with the consequent potential for misclassification would be avoided. (See Bramley & Dhawan, 2012 for a discussion of grade bandwidth and reliability).

A potential advantage is that there might be a positive effect on teaching. The current system 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. 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. With a 'fixed boundaries' scheme by definition there would be no difference in difficulty. One consequence would be that each raw mark gained in a unit would 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.

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 syllabus 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 at the targets 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 boundaries to move closer to the 'targets' 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

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

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).

Figure 3 below shows the grade A boundary marks over time for each of the units of the Mathematics A level whose structure was given in Figure 1. January sessions are shown in blue and June sessions shown in red. The black 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.

There are some interesting observations to be made from the pattern of boundaries in different units. There seem 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. Some of the units showed greater fluctuations in grade boundary than others, which is puzzling. 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. The latter 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.

Figure 4 below is a similar plot of the grade A boundaries for an A level Physics examination. The assessment structure for this examination (the equivalent of Figure 1) is given in the Appendix. There seem to be consistent Jan/June differences for unit G481, with higher boundaries in the June sessions than in the January ones. Also noteworthy is the rise in the boundaries on unit G485 in 2012, which was the result of a deliberate policy to make the paper easier and move the boundaries towards their target levels.

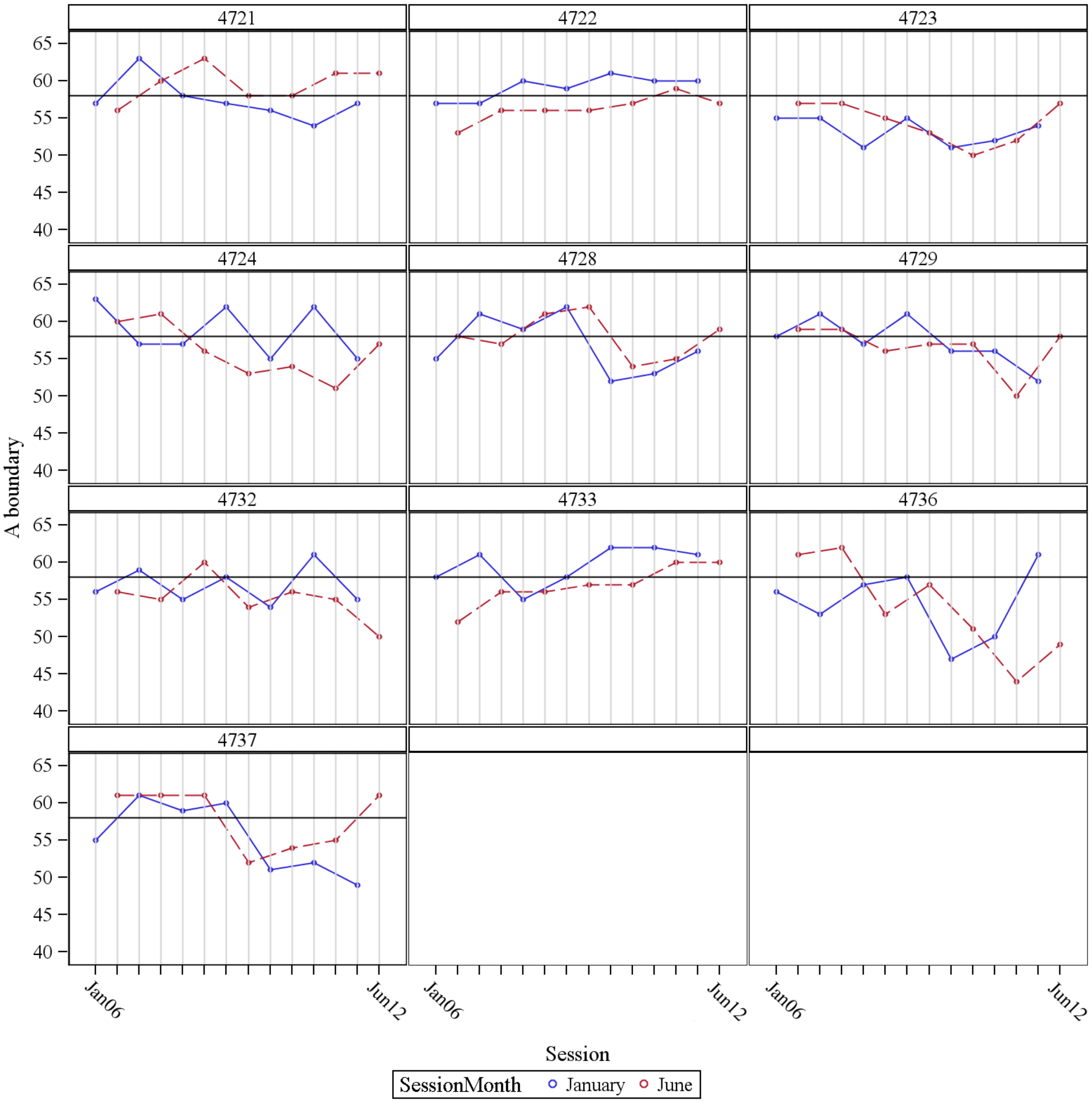

Figure 3: Location of the grade A boundary on the raw mark scale (range=0 to 72) in each unit of a Mathematics A level.

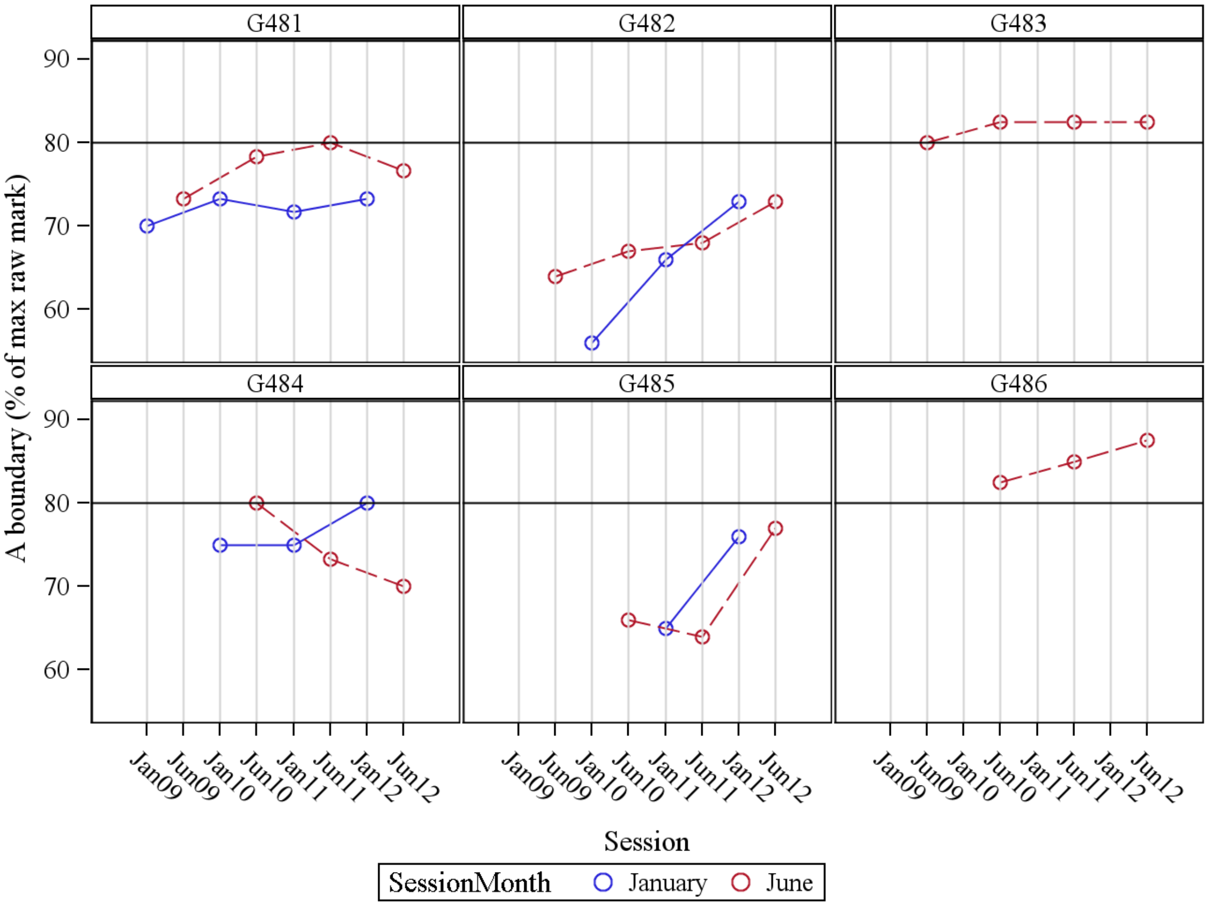

Figure 4: Location of the grade A boundary on the % raw mark scale (the units varied in maximum raw mark, see the appendix) in each unit of a Physics A level.

Effect of fixing the boundaries

In order to investigate the effect on the grade distributions of implementing a ‘fixed boundaries’ scheme, the UMS conversions were re-calculated for all the units in all the examination sessions shown in Figures 3 and 4, setting the unit grade boundaries at their ‘target’ values (represented by the horizontal black line in the graphs). The new unit UMS scores were then re-aggregated for the examinees who certificated in each of the June sessions, and hence the new grade distributions were found.

Of course, this kind of retrospective analysis cannot really answer the question of what would have happened in reality if these boundaries had been used, because examinee behaviour (particularly regarding re-sits) would have been different. Nonetheless, it does provide an indication of the likely impact on the grade distributions and more importantly an idea of how they would fluctuate over time.

Table 1: Maths A level – original and new cumulative percentage grade distributions[10]

	2009 (n≈11300)		2010 (n≈12500)		2011(n≈13000)		2012 (n≈13000)	
Grade	Original	New	Original	New	Original	New	Original	New
A	47.1	42.5	47.2	35.3	48.4	34.8	48.6	42.9
B	68.2	65.7	68.1	61.3	68.9	61.7	69.6	67.6
C	83.7	82.4	82.8	79.3	83.8	80.3	84.5	84.1
D	93.2	92.8	92.2	90.6	93.2	91.9	93.5	93.8
E	98.2	98.0	98.0	97.4	98.0	97.6	98.4	98.5

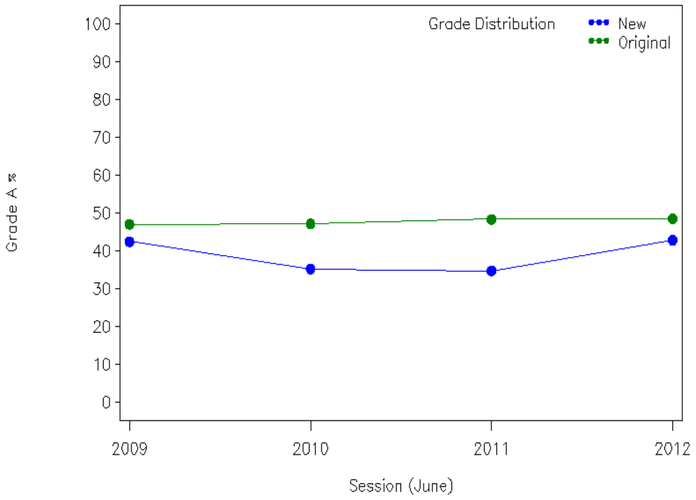

Figure 5: Maths A level – percentage of examinees gaining grade A over time.

Table 2: Physics A level – original and new cumulative percentage grade distributions.

	2010 (n≈6600)		2011 (n≈7700)		2012 (n≈8400)	
Grade	Original	New	Original	New	Original	New
A	29.9	8.6	30.4	9.4	29.6	18.1
B	50.4	33.8	52.0	37.4	52.8	51.9
C	69.0	60.9	70.6	65.3	72.0	78.1
D	84.4	83.5	86.0	87.4	86.9	93.8
E	95.9	97.0	96.4	98.3	96.6	99.3

[10] The A* grade, first introduced in 2010, is not considered here. Note that the 'original' figures will not exactly match published statistics because in practice the units selected for aggregation might have differed for examinees who also took the Further Mathematics examination. Also the datasets used may not have taken account of grade changes following result enquiries, appeals, or estimated grades given to examinees who missed units (e.g. because of illness).

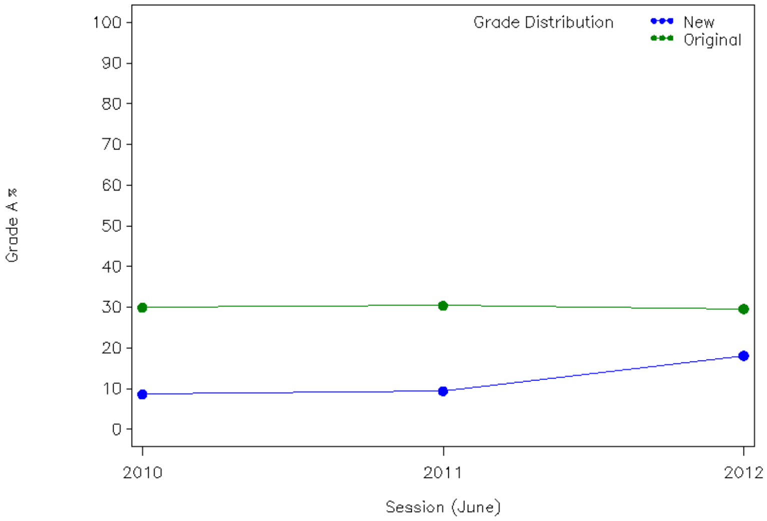

Figure 6: Physics A level – percentage of examinees gaining grade A over time.

It is clear from the graphs and tables that the effect of fixing the unit grade boundaries would have been to create much greater fluctuations over time in the cumulative percentage pass rates. In the Maths example, the largest year-on-year fluctuation at grade A would have been 8.1 percentage points (from 2011 to 2012), and in the Physics example it would have been 8.7 percentage points (also from 2011 to 2012). The corresponding largest fluctuations using the actual system were 1.2 and 0.8 percentage points. The differences are much greater at the higher grades – since most examinees achieve grade E there are fewer examinees at that point of the raw mark distributions on the units and hence changes in the boundaries have less effect.

It is also interesting to note that the effect of fixing the boundaries at the ‘target’ values would have been to reduce the cumulative percentage of examinees in each grade, as would be expected from consideration of Figures 3 and 4 where it is clear that the majority of unit grade A boundaries were below the target value of 80%. It may not have been necessary to introduce the A* grade if a ‘fixed boundaries’ system had been used.

Discussion

The above analysis has shown that grade boundaries on the units of two modular A levels have fluctuated considerably over time, but that the overall aggregate grade distributions have remained relatively constant. Fixing the unit grade boundaries at constant values would consequently have created much greater fluctuations from year to year in the grade distributions, particularly at the higher grades.

Such large fluctuations would probably not be acceptable to a public that has become accustomed, over a long period of time, to small (usually increasing) changes from one year to the next in the cumulative grade distributions. The (perfectly reasonable) assumption at the heart of standard maintaining processes has been that if there is no reason to believe that one year's cohort for an exam is different from the previous one, then there is no justification for a large fluctuation in grade distributions (an assumption dubbed by Newton (2011) as the 'similar cohort adage'). When exam boards have wanted to justify a larger-than-usual change in grade distributions, it has usually been by arguments that the cohort has changed in a way that warrants the change.

As the modular A level system bedded down in the years following 2002, the examination system achieved a reasonable balance of small fluctuations in grade distributions over time and comparability between AOs, while allowing schools and students the flexibility inherent in courses with a modular assessment structure. However, the general satisfaction with A level outcomes was obtained at the cost of a lack of transparency and general understanding of the two technical innovations that had made this possible, namely the Uniform Mark Scale and the 'comparable outcomes' (prediction matrices) method of standard maintaining.

But these innovations had not overcome the problem of the vulnerability of modular assessment structures when they are first implemented. The introduction of modular A levels that were first certificated in 2002 created great challenges for maintaining standards in the 'usual' way (see Baird, 2007 for some examples of problems that arose; also Tomlinson, 2002 for the report on the official enquiry into what had gone wrong). The main difficulty with modular systems is that the grade boundaries on units taken early in the course (e.g. in January) cannot later be changed (in June). Therefore exam boards wishing to obtain a particular grade distribution justified by the available evidence about the certificating cohort in June are only able to achieve this by changing boundaries on the units taken in June.

This constraint led to the recent crisis in the grading of GCSE English in June 2012, which gave rise to a legal challenge from schools and teaching unions and a judicial review (see House of Commons Education Committee, 2013). At the heart of the dispute was the belief that rises in the grade boundaries on coursework units between January and June in 2012 were unfair. Coursework units are marked by schools and moderated by AOs. Since the marking criteria do not change, there is some justification for the belief that the 'difficulty' of these units does not change either, which implies that grade boundaries should not change much, if at all, from one examination session to the next. (A similar argument could be made for other non-coursework units that consist mainly of long essay questions marked using mark schemes that apply generic criteria which do not change over time). This therefore has the potential to create a clash if the evidence is that the cohort is not very different from the previous cohort in terms of ability and therefore should obtain a similar distribution of grades, but when this can only be achieved by altering boundaries on coursework units.

The reason GCSE English was particularly affected was because of the relatively large weighting of coursework units in the total assessment, and the strong (and perhaps justifiable) belief by teachers that the difficulty of coursework units is indeed fixed. My suggestion is that conflicts of this nature arise more often than is commonly supposed within the A level and GCSE modular systems but go largely unnoticed because everyone is focused on the outcomes (grade distributions). Evidence supporting this suggestion can be found whenever there are inexplicable patterns in grade boundary locations, such as systematic differences between boundaries in January and June, as shown for some units in Figures 3 and 4.

I suggested (Bramley, 2012) that the benefits of fixing the boundaries at unit level in modular A level and GCSE assessments might outweigh the disadvantages, but accepted it would probably be too radical a change ever to be implemented in practice. In fact, both GCSEs and A levels in England are now undergoing radical reform – but the things that seem most likely to change are that the modular system will be replaced by the previous 'linear' system (where all assessments are taken in June at the end of the course); and that the amount of coursework will be substantially reduced! Further details about the planned changes can be found on the websites of the regulator Ofqual[11] and the Department for Education[12] . Although the main reasons for the reforms are not primarily to do with the technicalities of standard maintaining, these two changes will make it easier to apply the current 'comparable outcomes' approach. My hope is that we will not miss the opportunities for improved assessment design and increased transparency that would result if we had a greater focus on question and test difficulty.

Summary

In summary, conceptualising standards and standard-maintaining involves operationalising some apparently simple but actually very difficult concepts like 'test difficulty' and 'examinee cohort ability'. The standard-maintaining process used in England for GCSEs and A levels has always been very ability-focused – the main assumption being that, all things being equal, the distribution of grades in two large cohorts of similar examinees should not change much if at all from year to year. The introduction of modular assessment led to two innovations – the Uniform Mark Scale (UMS) which allowed aggregation of unit scores into assessment scores on a single scale for examinees taking different units at different times; and the 'comparable outcomes' (prediction matrices) method of standard maintaining. Whilst these innovations allowed modular assessments to run relatively smoothly once they bedded in, they introduced considerable complexity into the system which reduced public understanding, and, when problems arose, reduced public confidence.

The traditional focus on examinee cohort ability has been at the expense of a concern about test difficulty, with the result that within some modular assessments, implausible patterns of unit grade boundaries can be observed. If instead the underlying assumption was that the difficulty of assessment units should not change much if at all from year to year (implying fixed, or only slightly varying grade boundaries) there would be some benefits. In particular, there would be greater transparency to the examinees about what they needed to do to obtain each grade, and more straightforward inferences from the grade to the achievement on the test (assuming question papers and mark schemes are published). The main drawback to fixing the unit grade boundaries, as illustrated by the two examples in this paper, is that grade distributions would fluctuate far more than they currently do from year to year. This would probably be unacceptable to students, schools and to policy-makers.

However, we have seen that it can also be unacceptable when grade boundaries on coursework units are changed, and by extension it should also cause concern when boundaries fluctuate on exam papers that are marked according to generic criteria. These conflicts between what seem intuitively plausible assumptions about cohort ability and question paper difficulty partly explain why it is impossible to please everyone when attempting to maintain standards in examinations.

[11] http://ofqual.gov.uk/qualifications-and-assessments/qualification-reform/a-level-reform/ Accessed 13/08/13.

[12] http://www.education.gov.uk/schools/teachingandlearning/qualifications/gcses/a00221366/gcse-reform Accessed 13/08/13.

References

Acquah, D. K. (2013). An analysis of the GCE A* grade. *Curriculum Journal*, 1-24.

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

Benton, T., & Lin, Y. (2011). *Investigating the relationship between A level results and prior attainment at GCSE.* Coventry: Ofqual.

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, Potsdam, Germany, August 2008.

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

Bramley, T. (2012). *What if the grade boundaries on all A level examinations were set at a fixed proportion of the total mark?* Paper presented at the Maintaining Examination Standards seminar, London, March 2012.

Bramley, T., & Dhawan, V. (2012). Estimates of reliability of qualifications. In D. Opposs & Q. He (Eds.), *Ofqual's Reliability Compendium* (pp. 217-319). Coventry: Office of Qualifications and Examinations Regulation.

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

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.

House of Commons Education Committee (2013). 2012 GCSE English Results. First report of session 2013-14. http://www.publications.parliament.uk/pa/cm201314/cmselect/cmeduc/204/204.pdf Accessed 18/07/13.

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. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*: Cambridge University Press.

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.

Ofqual (2011). *GCSE, GCE, Principal Learning and Project Code of Practice.* Coventry: Ofqual. http://www2.ofqual.gov.uk/for-awarding-organisations/96-articles/247-code-of-practice-2011. Accessed 13/08/13.

Taylor (2012). *GCE predictions using mean GCSE score as the measure of prior attainment*. Report to Joint Council for Qualifications (JCQ) Standards and Technical Advisory Group (STAG), 23/03/13.

Thomson, D. G. (1992). *Grading modular curricula.* Cambridge: Midland Examining Group.

Tomlinson, M. (2002). Inquiry into A level standards. Final Report. London: Department for Education and Skills.
http://image.guardian.co.uk/sys-files/Education/documents/2002/12/03/alevelinquiry.pdf
Accessed 13/08/13.

Appendix

Unit code	Unit name	Type	Max Raw	Max uniform
G481	Mechanics	AS	60	90
G482	Electrons, Waves and Photons	AS	100	150
G483	Practical Skills in Physics 1	AS	40	60
G484	The Newtonian World	A2	60	90
G485	Fields, Particles and Frontiers of Physics	A2	100	150
G486	Practical Skills in Physics 2	A2	40	60

For a certificate candidates must have taken the above six mandatory units.

Figure A1: The structure of one particular 6-unit A level physics assessment.