CAMBRIDGE ASSESSMENT

On the limits of linking: experiences from England

Tom Bramley[1], Anthony Dawson[1] and Paul Newton[2]

Paper presented at the 76th annual meeting of the National Council on Measurement in Education (NCME), Philadelphia, PA, April 2-6, 2014.

[1]Cambridge Assessment, 1 Hills Road, Cambridge, UK, CB1 2EU.
[2]Department of Curriculum, Pedagogy and Assessment, Institute of Education, University of London, 20 Bedford Way, London, UK, WC1H 0A.

Introduction

Maintaining standards in public examinations in England raises some complex and possibly insoluble problems. Similar problems are doubtless faced in many jurisdictions and testing regimes, but the situation in England is arguably unique. The purpose of this paper is to show how the mechanisms for maintaining standards are stretched to or beyond their limit when there are significant changes in the examination system. This will be illustrated by consideration of two different crises that arose – the first in 2002 when the new 'Curriculum 2000' A levels were first awarded; and the second in 2012 when new GCSE qualifications in English were first awarded.

GCSEs and A levels are high-stakes examinations, generally in academic subjects, taken by students aged 16 (Year 11) and 18 (Year 13) respectively. Students usually take 8-10 GCSEs, and 3 or 4 A levels. GCSEs are reported on an 8-category letter scale from A* (high), A, B … to G (low), with U (unclassified) given to scores not worthy of a G. A levels until 2010 were reported on a 5 category scale from A to E with U given to scores not worthy of an E. Since 2010, an A* has also been available at A level.

In England, three Awarding Organisations (AOs) ('examination boards' would be a term with more global recognition) are accredited by the regulator (Ofqual) to provide GCSEs and A levels. They each offer a similar suite of GCSEs and A levels and, consequently, compete for their share of the qualifications market. To ensure that certificates from different AOs (e.g. GCSEs in Biology) convey a similar meaning, the regulator ensures that the syllabuses provided by each AO are broadly similar in terms of content and assessment structure. It is also responsible for ensuring comparability of standards, which in practice means ensuring two things:

- that a given grade from one AO 'means the same' as the same grade from another AO;
- that a given grade from each AO means the same as the same grade from that AO in previous years.

A third possible kind of comparability – across different subjects (e.g. that an A in Chemistry 'means the same' as an A in French), is not explicitly regulated, although attention is given to ensuring broad alignment of similar subjects, like Biology, Chemistry and Physics; or French, German and Spanish.

Thus 'content standards' are maintained by ensuring that each AO constructs its assessments such that they cover the assessment objectives set out in its syllabus, the content of the latter having been approved by the regulator. There are no explicit 'assessment difficulty standards' that the AOs must meet, in that there is no requirement for the cut-scores representing achievement at the different letter grades to be at the same percentage points on the raw score scales[1]. However, great attention is given to ensuring equivalent 'performance standards' in the sense that the letter grades are supposed to represent comparable achievement.

The process by which the cut-scores for the letter grades (known as the 'grade boundaries') are established is known in England as 'awarding'. The procedures for awarding are complex and detailed. They are prescribed in the regulator's Code of Practice (e.g. Ofqual, 2011). A large amount of research has been devoted over the years to investigating and clarifying issues such as: What is meant by 'standards' (e.g. Baird, et al, 2000; Bramley, 2005; Cresswell, 1996; Christie and Forrest, 1981; Massey, 1994); what is meant by 'comparability' (e.g. Bardell, et al, 1978; Forrest and Shoesmith, 1985; Coe, 2010; Newton, 2010a,b; Elliott, 2011); what sort of evidence (judgmental or statistical) should be given more weight (e.g. Cresswell, 2000); what the strengths and weaknesses of different statistical and judgmental methods are (e.g. Newton, et al, 2007). A further distinction of operational significance is between methods that can be applied in real time to inform the setting of 'live' grade boundaries, and methods that can only be applied post hoc to investigate whether comparability was achieved (in the sense relevant to those methods).

[1] Although there is a general aim for the grade A and E boundaries on each unit to be at roughly 80% and 40% respectively of the raw mark total in order to ensure a smooth linear transformation of raw scores to UMS scores: see Appendix A.

The procedures are placed under particular strain whenever there is significant change in the system. Such change could be the replacement of an old syllabus by a new syllabus, or a modification to the assessment structure. Changes in syllabus content can lead to charges of 'dumbing down' if it appears that the content removed has been replaced by less content, or less demanding content. But how should performance standards be maintained? If we assume that in the first year of a new syllabus, the unfamiliarity of teachers with the new content and with the new assessment structure leads to examination performances that are on average worse, should the awarding procedures allow for this in some way? And if in subsequent years there is a genuine improvement in performance, should this be reflected in better grades? Increases in the proportion of examinees receiving higher grades are also taken as evidence of 'grade inflation' and 'dumbing down' (e.g. Stewart, 2009). A mechanism by which syllabus change could lead to grade inflation was first described by Pollitt (1998); it resembled a similar mechanism which had been discussed in the North American context by Linn et al (1990).

The 2002 crisis

The first 'crisis' we describe is that of the grading in 2002 of the new A levels courses which had been taught for the first time in September 2000. All AOs had developed new syllabuses with substantial modifications in content, reflecting the need to keep them up-to-date in terms of new knowledge and changing values. The most significant change, however, was in assessment structure. Prior to 2002, the majority of A levels had been 'linear' exams, meaning that all the examination took place at the end of the two-year course. From 2000, the syllabuses prescribed 'modular' or 'unitised' assessment – the assessments being split into discrete modules or units which could be taken (subject to syllabus-specific restrictions) at any of the four examination sessions between the start of the course and the end of the course, namely those in January and June of the first year of the course, and those in January and June of the second year of the course. A designated half of the units, intended (but not required) to be taught and examined first, would count towards a separate qualification, the AS level. The educational point of this was to allow students to study a broader curriculum in the first year of the A level course, perhaps taking 5 AS levels, and then in the second year specialise to 3 A levels, with the units from the AS level being supplemented by units designated as 'A2'. The A2 units did not form a separate qualification, but in combination with the AS units led to an A level.

One problem created by this change in structure was how to 'split' the standard enshrined by the linear A level into a standard for the AS qualification and one for the new modular A level. The solution was to have the AS units graded at a slightly lower standard than the linear A level, to reflect candidates' educational development and maturity during the first year of an A level programme; and the A2 units at a slightly higher standard, to reflect candidates' developing capacity during the second (Tomlinson, 2002a). Thus an examinee who got a B at AS level would be expected to get a B at A level a year later, after a year's worth of consolidation and maturation.

Another major problem was how to fairly aggregate results from different units, given that examinees in a given cohort could have taken different combinations of units, or the same combination of units taken in different patterns of examination sessions (completely different exam papers were created for each new session). The solution to this problem was the introduction of the Uniform Mark Scale (UMS), essentially a finer-grained version of the grade scale. Appendix A contains a description of how this scale works.

The specific problem created in 2001 and 2002 was how to set the grade boundaries on the AS units and January 2002 A2 units in such a way that when the boundaries were set for the A2 units in June 2002, the main cohort of 'aggregating[2]' examinees would obtain an A level grade distribution which was in some sense comparable to those of previous (linear A level) cohorts – in other words that it would be possible to maintain the standard at the overall qualification level, in some sense, without jeopardising them at unit level. The sense in which standards were to be

[2] Examinees can aggregate their unit scores in any session where they have a valid combination of units to aggregate. The most typical route involves students taking their final unit(s) at the end of the two-year course and aggregating all their units at this point to obtain the A level.

'maintained' at the overall qualification level was described as the 'comparable outcomes' perspective: "candidates taking the new exams should receive, as a group, comparable grades to those which they would have received had they followed the old courses." (Cresswell, 2003, p.14). The fact that in June 2002 many examinees had already received grades and UMS scores on many of their units (AS and some A2) meant that the AOs were placed in the position of having to locate new unit boundaries in what could have been unusual, unforeseen or at worst indefensible places on the June 2002 A2 units, given the need to ensure similar overall grade profiles to those awarded from corresponding linear A levels. It was therefore an 'accident waiting to happen'.

For certain A2 units, a clear tension arose between grade boundaries that would be required by the comparable outcomes perspective and the standard of performance that subject matter experts (examiners who comprised the awarding committees) were seeing at those boundaries. This led to conflict between examiners who prioritised evidence of performance in scripts (comparable performances) and board officers who prioritised evidence of statistical impacts (comparable outcomes). Consequently, for some units on some examinations, the 'accountable officer[3]' at the AO made a decision to raise the grade boundary after it had been duly set by the expert panel at the awarding meeting. This proved particularly problematic in assessments where the structure allowed examinees a choice between a written unit and a coursework unit. The former are traditional time limited externally marked examination papers, whereas the latter are open-ended tasks set by the examinee's school and marked (scored) by their teachers to generic criteria specified by the AO. A sample of each school's marks is externally 'moderated' by the AO to check that schools are applying the generic mark schemes in the same way. Because of the generic marking criteria, the grade boundaries on coursework units are not expected to change from one session to the next. With such units the general pattern is for boundaries to stay in the same place for several sessions and then to rise by one mark if necessary to control overall grade inflation. There is also a tendency for examinees to obtain better grades on coursework units than on written units. In 2002, in order to control grade inflation and align coursework grade distributions better with their corresponding written papers, in some instances board officers had changed boundaries on coursework units such that the A and E boundaries ended up unusually close together, with the result that some examinees (especially in the fee-paying sector) ended up with bizarre-looking profiles of grades across their units, such as AAAAAE. The ensuing crisis created headlines such as 'Exam board admits lowering grades'[4]; and 'A-level scandal is not only incompetence, but also deceit'[5].

An official inquiry was launched by the Secretary of State for Education and Skills into the grading of the summer 2002 exams. It had two phases. The first phase specifically investigated allegations of impropriety in the setting of standards (grade boundaries) on units in the June 2002 exam session. Its findings were reported in Tomlinson (2002a), and led to the review of grades awarded in one or more units in 31 separate A level subjects. 9,800 unit grades were improved, resulting in an improvement in overall grades for 1,945 examinees. This action was taken because he (Tomlinson) considered that the accountable officers of the AOs had felt that that they had been put under undue pressure from the regulator (at the time known as the Qualifications and Curriculum Authority, QCA) to produce comparable outcomes, despite the quality of performances evident at grade boundary scripts not matching the expectations of subject matter experts. In other words, boundaries had been moved solely to produce what the AOs felt was the 'expected' distribution of grades. Tomlinson was clear that this did not arise from impropriety or lack of integrity at the AOs; but rather from a lack of guidance in the procedures on the relative weight to give to judgments about the quality of work on the one hand and statistical information on the other.

The aim of the second phase was "to investigate the arrangements … for setting, maintaining and judging A level standards … and ensuring their consistency over time; and to make recommendations … with the aim of securing the credibility and integrity of these exams."

[3] Designated person (in many cases the Chief Executive) responsible for all grading outcomes in an AO.
[4] http://www.theguardian.com/uk/2002/sep/18/alevels2002.schools
[5] http://www.telegraph.co.uk/comment/personal-view/3581778/A-level-scandal-is-not-only-incompetence-but-also-deceit.html

(Tomlinson, 2002b, p8). It is clear from this report that Tomlinson did not endorse the above definition of standard maintaining (that similar groups of examinees should obtain a similar distribution of grades), as the following quotes show:

> "I must stress that the use of statistical and other information in this way is a wholly legitimate and necessary part of the process of maintaining standards. But it is clear that part of the problem in 2002 was the perception that too much weight was accorded to statistical information in the grading process for some subjects, and particularly in modifications made in the latter stages of the awarding process to the grade boundaries recommended to awarding body Accountable Officers by their Chairs of Examiners. It is in my view essential to restore confidence in the grading process as a whole, and particularly the legitimate use of statistical information." (Tomlinson, 2002b, p22-23).

> "But longer term action is needed to ensure that similar concerns do not arise in 2003 and subsequently. QCA, together with the other regulatory bodies, is currently revising the Code of Practice governing A level grading in time for the January 2003 examinations to: *give greater emphasis to examiners' judgements about the quality of candidates' work* and to clarify the role for statistical evidence in awarding; to ensure consultation between Accountable Officers and Chairs of Examiners before grade boundaries are finalised; and to ensure that QCA can monitor grade boundary decisions made by Accountable Officers. These are important steps towards improving the transparency and consistency of the awarding process." (Tomlinson, 2002b, p22-23, *italics added.*)

Thus this crisis brought to the fore the potential conflict between statistical information (for example about cohort ability and potential pass rates) and the judgment of expert examiners about the quality of examinees' work. Such a conflict potentially exists each year, of course, where experts are expected to allow for differences in the difficulty of the current examination when judging the quality of work produced. However, in this case the experts were supposed somehow also to allow in their judgments for differences in content standards and assessment structure between the old and new syllabus while still being able to recognise the same standard of achievement.

How *should* the experts have allowed for the effect of moving from a linear to a modular examination? It is not clear whether this was ever explicitly stated. One analogy[6] is comparing performances (running times) in a 400m athletics race with running times in an event where athletes have to run four separate 100m races and their running times for each are added together. The implication from this analogy is that modular exams are 'easier' in the sense that just as it is presumably easier to get a faster time for 400m if you can run it in stages, so it is easier to produce a good aggregate examination performance if you can study and revise for one unit at a time. Extending the analogy to allowing re-sits of individual units would correspond to athletes being allowed to re-run particular 100m sections and use their fastest time to count towards their 400m total time, again implying that modular examinations are easier.

However, one of the motivations for the introduction of modular courses in the first place was that it would allow for better integration of learning and assessment, put less emphasis on last minute 'cramming', reduce test-taking stress, and so on. If these points are accepted there is an implication that examinees' performances in a modular assessment will be genuinely better. Indeed, one of the post hoc justifications for allowing grade distributions to rise during 2002 was that the modular approach had provided students with additional information on their strengths and weaknesses – bearing in mind that the new approach encouraged students to take a broad range of subjects in year 1 and to specialise in their 'stronger' subjects during year 2 – almost implying that overall A level performance ought to improve under the new approach (Baird, et al, 2003). The question for AOs (and their stakeholders) is whether it is right for pass-rates to rise to reflect this improvement in performance, and what evidence could possibly be gathered to show that the improvement is real rather than imagined.

Set against this, however, is the point mentioned above, that when an assessment structure and syllabus content change, the unfamiliarity puts examinees at a disadvantage, at least for the first

[6] Seen in a letter to the Times Educational Supplement.

year's cohort and possibly for subsequent cohorts, while teachers get used to teaching the new course. Therefore one might perhaps expect (imagining for a moment that it is possible to identify a genuine standard of achievement) pass-rates to fall in the first year after the change from linear to modular assessment, but then rise over a period of time to a level above that of the old linear assessment.

Furthermore, returning to the athletics analogy, while it is presumably true that running times on the 4x100m event would be lower than in the 400m event in absolute terms, it is also true that in relative terms, some athletes would benefit and some would be disadvantaged, in terms of where they would appear in the overall rank order if the same group of athletes had to do both. To spell it out, good sprinters would presumably do relatively better in the 'modular' race whereas those with more endurance would presumably do better in the single longer race. That is, the change of assessment structure changes the construct being assessed. The greater the change, the less sense it makes to claim that there is a standard of achievement that can be carried from one to the other, if this standard is defined in terms of KSUs (knowledge, skills and understanding).

Probably for this reason, and in contrast to the view expressed above by Tomlinson, a consensus seemed to emerge in the 2000s that the appropriate way to maintain standards when assessment syllabuses and structures change is indeed to try to ensure similar grade distributions for similar groups of examinees. As noted earlier, this approach – effectively a definition of standard maintaining – came to be known as the 'comparable outcomes' approach (Cresswell, 2003). The main issue is how to determine how similar groups of examinees are, and then how to adjust grade distributions to reflect any dissimilarity. One crude measure of similarity that had been used by the AOs in the 1990s had been the proportion of examinees coming from different types of school, on the assumption that certain types of school (selective schools and fee-paying schools) tend to produce better examinees (see Eason, 1995). In the 2000s, thanks to the improved national collation of large longitudinal datasets, it became possible to use prior attainment as the 'common yardstick' to measure similarity.

The output of the comparable outcomes approach is a target or 'putative' grade distribution for the subset of examinees that has been successfully matched with a measure of prior attainment. At A level this measure of prior attainment is the mean GCSE score; and at GCSE the measure is based on Key Stage 2 (KS2) performance.[7] Each examinee is assigned to a decile (GCSE) or octile (KS2) based on the performance of the entire GCSE or KS2 cohort.[8] The cross-tabulation of prior attainment category and examination grade in year x forms what is known as an 'outcome matrix' where the cells contain the cumulative percentage within a prior attainment category obtaining each exam grade. The same percentages are applied to the new prior attainment distribution in year x+1 to produce a 'prediction matrix' for year x+1. Summing the columns of this matrix gives the putative grade distribution for year x+1. Bramley & Vidal Rodeiro (2014) have shown that this method is structurally very similar to the frequency estimation equipercentile method of test equating used with a non-equivalent groups anchor test (NEAT) design, although with some obvious major differences – not least the fact that the measure of prior attainment is not an anchor test and is taken several years before the exams being 'equated'.

The comparable outcomes (CO) method was used first for A levels in 2002. Initially it was just one more source of evidence for the AOs to take account of when setting the grade boundaries on examinations, both in 'normal' circumstances when there is no change of syllabus or assessment structure as well as at times of significant change. However, over time it came to assume greater importance and since 2010 (for A levels) and 2011 (for GCSEs) the AOs have had to ensure that the grade boundaries on examination units are set in such a way that the overall aggregate grade distribution falls within specified tolerances of the putative grade distribution. Any larger-than-tolerance deviations at particular grade boundaries need to be justified to the regulator.

[7] KS2 tests are taken by 11 year olds England in Maths, English and Science..

[8] This is a simplification – for fuller details see Benton & Lin (2011) and Taylor (2013)

The 2012 crisis

Following a large-scale curriculum and examination reform initiative, new modular GCSE syllabuses were introduced for first teaching in 2009 and 2010. The changes in content and structure were particularly significant for the new GCSEs in English, which were taught from 2010, with the first aggregation of unit results in June 2012. Once again, the comparable outcomes approach was used to derive putative grade distributions, and the boards duly attempted to set grade boundaries on the June 2012 units that would allow a reasonable fit to these putative distributions. Did this reassure the public that standards were maintained properly? Not in the least, if the following press headlines are anything to go by:

'English GCSEs marked down to curb grade inflation, say teachers'[9]
'Tinkering with GCSE outcomes hugely unfair on students'[10]

Once again, the media blamed the regulator as well as the AOs, with most reports making it clear that the AOs were under pressure from the regulator to move grade boundaries. A particular exacerbating feature of the English case is that the units on which the grade boundaries were changed included 'controlled assessment'[11] units where it might be thought that there was no difference in difficulty – the reason being that they were marked (scored) by a generic rubric that did not change from January to June. Hence there is less a priori justification for having grade boundaries that differ from one session to the next on these units. As noted previously, with coursework units the general pattern is for boundaries to stay in the same place and only to rise by one mark at a time. Here, in contrast, the boundaries on one particular controlled assessment unit rose by 10 marks out of 96 from January to June. It is not surprising that this was widely perceived to be unfair.

A legal challenge against two AOs and the regulator Ofqual was mounted, with local authorities, schools, teachers and pupils included among the claimants. The judgment on this case (London Borough of Lewisham and others v AQA, Edexcel, Ofqual and others, 2013) set out the problem with great clarity:

> The claimants' complaint is that too rigorous a standard was adopted when assessing some of the units in June 2012 with the result that many pupils who confidently and reasonably expected to attain the C grade, on the basis of results which their fellow examinees had obtained in the January 2012 and indeed earlier assessments, inexplicably failed to do so. There was an unheralded and unjustified shift in the grade C boundary. This constituted an elementary unfairness because pupils competing in the same examination were not treated equally. The January cohort of students was graded more leniently than the June cohort, at least in some of the papers assessed by the two AOs. Ofqual, as the regulator, had power to forbid this inconsistent and unfair treatment by issuing statutory directions, and its failure to do so in order to remedy this conspicuous unfairness constituted an error of law.
>
> This unfairness was, say the claimants, compounded by two further factors. First, both the AOs and Ofqual had led the pupils and their teachers to understand that the marking standard would be consistent at whatever stage in the two year cycle a unit was completed. The natural inference from this was that in relation to any particular unit, the same, or at least substantially the same, grade boundary would be adopted in June as in the previous January. It is conceded that everyone understood that there might be some minor variation in the mark boundary for written examination papers to reflect the fact that a particular paper may vary in difficulty from one half-yearly assessment to the next. The marks will then be correspondingly higher or lower depending upon whether the paper is easier or harder and the grade boundary will need to be adjusted accordingly, but no radical change would have been anticipated in such cases. For controlled assessments, where the task remains precisely the same whenever the unit is completed, there is no justification in changing the grade boundary at all. Mr Sheldon QC, counsel for the claimants, submits - and this is not disputed - that many pupils and teachers had acted on that assumption to their detriment. In

[9] http://www.guardian.co.uk/education/2012/aug/22/english-gcses-marked-down-teachers
[10] http://www.ascl.org.uk/news-and-views/news_news-detail.tinkering-with-gcse-outcomes-hugely-unfair-on-students.html
[11] Controlled assessment was introduced at GCSE in 2007-8. It was designed to reduce (amongst other things) the perceived opportunities for malpractice (e.g. parental assistance) possible with coursework. For GCSE English the controlled assessment tasks for reading and writing units were set by the AOs, but some choice of tasks by the school was allowed, along with opportunity to 'contextualise' the task. As with coursework, they were internally marked and externally moderated.

> some cases, for example, there is evidence that once teachers were confident that a student would achieve a C grade on the basis of previous grade boundaries, the student was encouraged to switch focus to other subjects. (Paragraphs 9 and 10).

What was less widely perceived was that it was the few January examinees who had been advantaged, rather than the large majority of June examinees who were disadvantaged. The judgment found in favour of the AOs and the regulator, who were seen to have used the best evidence available in January to set the grade boundaries on these units, but that this evidence was simply insufficient to allow appropriate boundaries to be set.

> In this case I am satisfied that the examiners in June made assessments which they thought fairly reflected the standard of the scripts. In the light of the fuller information then available to them, their judgments were more accurate and more reliable than the January assessments. Wider concerns about creating unfairness as between those qualifying in different years, and the need to retain the value of the qualification, strongly militated against applying the January grades to the June assessments (even with such modification as may have been necessary to account for more lenient marking) to the June assessments. There was no obligation to extend the generosity of January to June; on the contrary, there was every reason to correct the earlier erroneous standard. There was no unfairness, conspicuous or otherwise, in what they did. (Paragraph 129).

> The problem lies in the modular nature of the examination, coupled with the fact that grade boundaries were assessed and made public at each stage of the process. … having now reviewed the evidence in detail, I am satisfied that it was indeed the structure of the qualification itself which is the source of such unfairness as has been demonstrated in this case, and not any unlawful action by either Ofqual or the AOs. (Paragraphs 152 and 157).

Discussion

We have shown that two widely publicised 'crises' in the examination system in England, separated by 10 years and both widely reported in the media at the time, essentially arose from the same set of circumstances:

- A change of content standards and (especially) assessment structure;
- Modular examinations where decisions about grade boundaries taken in earlier sessions could not be altered;
- Units where the nature of the particular assessment would lead one to expect little or no change in grade boundary from one session to another;
- A perception, or correct understanding, from AOs that the regulator would not be satisfied with an unjustifiable sudden rise in pass rates.
- Conflict between examiners (subject matter experts) who felt that their expertise was being overruled to ensure a statistical 'fix'.

Both crises were investigated by external and independent authorities – in 2002 a public enquiry and in 2012 a judicial review. However, the two independent authorities reached opposite conclusions. In 2002 Tomlinson concluded that too much weight had been given to statistical considerations (comparable outcomes) and not enough to expert judgment. Some examination units were re-graded. In 2012 the judges fully supported the comparable outcomes approach and no units were re-graded. Both cases led to a distinct loss of public confidence in the public examination system.

Who was right? Of course, there is no answer to this question. It all depends on the choice of definition of what is being maintained when standards are maintained. If we think that it is a level of achievement in terms of knowledge, skills and understanding, recognisable by subject experts from consideration of question papers, mark schemes and examination performances, then a standard-maintaining method that gives due amount of weight to expert judgment will be appropriate. If we think that it is the proportion of examinees of similar ability (for example as defined by prior attainment) obtaining each grade, then a statistical method like the comparable outcomes method will be appropriate. If we wish to blur our definition so that it contains an element of both approaches then this might produce an acceptable compromise. There would still

be unresolved questions about how much weight to give to each approach. We have argued in several places (e.g. Black & Bramley, 2008) that the best compromise approach would involve sources of evidence that are independent of each other. The current procedures do not meet this ideal because the range of scripts (examples of examinee performance) considered by the subject experts is determined by prior consideration of statistical information.

Regarding the specific question of what to do when there is a significant change in the system, the earlier discussion of modular assessment suggests that it is probably too much to expect experts to judge performance standards in the light of changing content standards, assessment structures and the construct being assessed. In these cases a comparable outcomes approach seems reasonable. Is it then unreasonable to continue to use comparable outcomes for 'routine' standard maintaining when things are not changing? The answer to this question depends on the purpose of the qualification for the 'user' – and different users have different purposes. The comparable outcomes approach effectively ends 'grade inflation' by keeping the overall grade distribution (across all AOs) roughly constant – or, more accurately, the grade distribution for examinees with similar prior attainment constant. If the purpose (e.g. for a government) is to use GCSE or A level results to monitor absolute changes in attainment in the system then this will not be deemed satisfactory. Schools will not be happy with this approach if their own 'improvement' is defined in terms of absolute targets for proportions of students achieving certain grades at GCSE or A level. The students themselves might be reasonably happy in the sense that their grades will not change value too much from year to year. Teachers might be happy in the sense that their individual students can certainly benefit from improved teaching by them, provided that it is larger than any average global improvement in teaching. The comparable outcomes approach should also (but in practice does not always) reassure users that the competition between AOs for market share does not produce a 'race to the bottom' whereby the AOs compete by lowering their standards to have the qualifications in which it is easiest to obtain a given grade.

In conclusion, we have shown that the conflict between different definitions of standard maintaining is an ongoing source of tension even in 'normal' times when syllabuses and assessment structures are not changing. In times of change, it seems reasonable to adopt a comparable outcomes definition, but this then raises the question of what to do in the years following the change. If a performance-based definition is subsequently allowed to play a significant part, there is the possibility for a 'ratchet effect' and the risk of grade inflation. On the other hand, if a performance-based definition has a small or no part in standard-maintaining procedures there is the risk that the outcomes are seen as a 'statistical fix' with an attendant fall in stakeholder confidence.

References

Acquah, D. K. (2013). An analysis of the GCE A* grade. *Curriculum Journal, 24*(4), 529-552.

Baird, J., Cresswell, M. and Newton, P. (2000). Would the real gold standard please step forward? *Research Papers in Education, 15 (2)*, 213-229.

Baird, J., Ebner, K and Pinot de Moira, A. (2003). *Student choice of study in Curriculum 2000.* Guildford, Surrey: Assessment and Qualifications Alliance.

Bardell G.S., Forrest G.M. & Shoesmith D.J. (1978). *Comparability in GCE: A Review of The Boards' Studies, 1964-1977*. JMB on behalf of the GCE Examining Boards. Manchester.

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

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

Bramley, T. (2005). Accessibility, easiness and standards. *Educational Research, 47 (2)*, 251-261.

Bramley, T. (2013). *Maintaining standards in public examinations: why it is impossible to please everyone.* Paper presented at the 15th biennial conference of the European Association for Research in Learning and Instruction (EARLI), Munich, Germany, 27-13 August 2013.

Bramley, T. & Vidal Rodeiro, C. (2014). *Using statistical equating for standard maintaining.* Cambridge Assessment report.

Christie, T. & Forrest, G.M. (1981). *Defining public examination standards.* Schools Council Research Studies. Macmillan Education.
Coe, R. (2010). Understanding comparability of examination standards. *Research Papers in Education, 25*(3), 271-284.
Cresswell, M.J. (1996) Defining, setting and maintaining standards in curriculum-embedded examinations: judgemental and statistical approaches. In Goldstein, H. and Lewis, T. (Eds) *Assessment: problems, developments and statistical issues.* Chichester: John Wiley and Sons Ltd.
Cresswell, M.J. (2000). The role of public examinations in defining and monitoring standards. In H. Goldstein and A. Heath (Eds.) *Educational Standards.* Oxford: Oxford University Press.
Cresswell, M.J. (2003). *Heaps, prototypes and ethics: the consequences of using judgements of student performance to set examination standards in a time of change.* London: University of London Institute of Education.
Eason, S. (1995). *A review of the Delta Analysis method for comparing subject grade distributions across examining boards.* Guildford: Associated Examining Board.
Elliott, G. (2011). A guide to comparability terminology and methods. *Research Matters: A Cambridge Assessment Publication, Special Issue 2: comparability*, 9-19.
Forrest G.M. & Shoesmith D.J. (1985) *A second review of GCE comparability studies.* JMB on behalf of the GCE Examining Boards. Manchester.
Linn, R.L., Graue, E. and Sanders, N.M. (1990). Comparing state and district test results to national norms: the validity of claims that "everyone is above average". *Educational Measurement: Issues and Practice, 9 (3)*, 5-14.
London Borough of Lewisham & Ors v Assessment and Qualifications Alliance ("AQA"), Pearson Education Limited ("Edexcel"), Office of Qualifications and Examinations Regulation ("Ofqual") [2013] EWHC 211, Case Nos. CO/11409/2012 and CO/11413/2012
Massey, A.J. (1994). Standards are slippery! *British Journal of Curriculum and Assessment, 5,* 37-8.
Newton, P.E. (2010a). Conceptualizing comparability. *Measurement: Interdisciplinary Research and Perspectives, 8(4)*, 172-179.
Newton, P.E. (2010b). Contrasting conceptions of comparability. *Research Papers in Education, 25(3)*, 285-292.
Newton, P.E., Baird, J., Goldstein, H., Patrick, H. & Tymms, P. (2007). (Eds.). *Techniques for monitoring the comparability of examination standards.* London: Qualifications and Curriculum Authority.
Ofqual (2011). *GCSE, GCE, Principal Learning and Project Code of Practice.* Coventry: Ofqual.
Pollitt, A. (1998). *Maintaining standards in changing times.* Presented at the 24th Annual Conference of the International Association for Educational Assessment. Barbados.
Stewart, W. (2009). Exams: A-level results continue to rise. Published in TES Newspaper on 21 August, 2009. Online version (accessed 10 March 2014): http://www.tes.co.uk/article.aspx?storycode=6020934
Taylor, M. (2013). *GCSE predictions using mean Key Stage 2 level as the measure of prior attainment.* Report to Joint Council on Qualifications (JCQ) Standards and Technical Advisory Group (STAG). Revised 26/06/13.
Tomlinson, M. (2002a). Inquiry into A level standards. Interim report.
Tomlinson, M. (2002b). Inquiry into A level standards. Final report.

Appendix A – Description of the Uniform Mark Scale (UMS)

The description below is taken from Bramley (2013).

> 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[12]. 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, for a unit where the maximum UMS is 100 and the maximum raw mark is 72*], 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[13]. 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%. […]. 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. (Bramley 2013, p4, italicized text adapted.).

[12] 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.

[13] 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.