English Baccalaureate Certificate (EBC) proposals: examining with and without tiers

John Hamer, Professor Roger Murphy, Dr Tom Mitchell, Dr Anna Grant and Jenny Smith

Contents

1. Introduction

The English Baccalaureate Certificate (EBC) consultation paper has heralded the prospect of some significant moves away from the pattern of examining which over a period of 25 years has become the established GCSE approach to examining the full range of students reaching the end of Key Stage 4 in England, Wales and Northern Ireland. The Secretary of State is known to be critical of GCSEs, and since coming to office has referred to them as constituting part of a 'discredited examination system'. At the heart of his criticisms has been a conviction that GCSEs have become too easy, and have become part of what he sees as 'a race to the bottom'. This expression is intended to portray the improvements in achievement in relation to the top GCSE grades as resulting from the Awarding Organisations, teachers and students colluding together to make sure that this happened through a range of dubious activities. These include: repeated entries; easier questions; giving out information about the content of future exams through textbooks and training courses; the lowering of grade boundaries; generous teacher assessments; and 'teaching to the test'. Among the proposed steps to improve things through the EBC developments it is proposed that teacher assessment will be abolished, the core EBC subjects will each only be examined by one Awarding Organisation, and every EBC subject will be assessed by a single external examination, which ideally will include no tiering to accommodate the needs of students with different levels of achievement. The consultation document asserted that:

> *There is currently a GCSE system with two tiers of exams, which caps students' aspiration. Students are able to take either foundation or higher tier examinations in a number of GCSEs, including all of the English Baccalaureate subjects except history. The higher tier allows students to achieve grades A* - D and the lower tier allows students to achieve grades C - G. In ordinary circumstances, it is rare for a student who is entered for the foundation tier to achieve higher than a C grade. The prospects for those students taking a foundation tier paper are poor; progression rates for students achieving C grade are much lower than for those achieving A* - B. Further education institutions frequently require a B grade or higher at GCSE for access to some A Level courses. Having a grade-cap in foundation tier examinations is also likely to be de-motivating and limit the aspirations of students.*[1]

On this basis the consultation document went on to point to the way ahead as being to avoid tiering 'wherever possible'.

> *'We believe that wherever possible EBCs should not be tiered, removing the grade-cap that currently exists at C grade in GCSE foundation tier papers, to benefit all students and increase motivation and attainment possibilities.'*[2]

In this paper we consider this development in the context of the considerable body of research evidence about the advantages and disadvantages of examining a wide range of students, with varying levels of achievement, using tiered and non-tiered examinations. In doing so we are concerned to review not just what is possible in terms of setting appropriate examination papers of both types, but also:

[1] Department for Education (2012), Reforming Key Stage 4 Qualifications, Consultation Document, September 2012, para 5.7
[2] Ibid, para 5.8

- to examine such evidence as exists about the impact of tiering or non-tiering on students' attitudes, aspirations and achievements; and
- to reflect upon the impact on classroom practices and overall achievement levels.

The obvious place to start is to take a close look at the advent of tiering, and the reason that it has been a central part of GCSE examining for over 25 years. Later in this paper will look at evidence from particular subjects and then broaden our review to experiences gained in different countries.

2. Background: the development of GCSE

The original calls from teachers for a common examination at 16-plus – now nearly half a century ago – envisaged an examining system within which all students who were entered followed the same syllabus and took the same examination. The examination boards, test developers and the Schools Council, however, were conscious of the technical difficulties in examining over what was felt to be such a wide range of ability.[3] Accordingly, the 1975 Schools Council document referred to a common 'examining system' by which was meant one in which there was **either**:

- a choice of questions/papers of different levels of ability with limits on the range of grades that might be gained by different routes: **or**
- where a choice of content or method of assessment was possible, but with the full range of grades being available whatever the option(s) chosen.[4]

The first reference to 'differentiated papers' came three years later in the report of the Waddell committee. They reported in favour a single system of examining, but recommended the introduction of a range of papers designed to cater for the extremes of the ability range. After reviewing the various trial examinations that had been undertaken, the report concluded that whereas the common approach was successful in some subjects, in others it was less so. This was particularly the case with mathematics and modern languages.[5]

The Waddell Report identified a range of five examining strategies, the first three of which it noted were usually referred to as 'common examinations':

> *(i) common papers taken by all candidates;*
>
> *(ii) common papers taken by all candidates, but containing questions designed to present different degrees of difficulty (for example, structured questions which all candidates are expected to attempt, and which have a built-in incline of difficulty);*

[3] Gipps, C. (1986). *GCSE: Some Background*, in *The GCSE: an uncommon examination*, Bedford Way Papers, No 29, Institute of Education, University of London, 1986; and *A Common System of Examining at 16-plus,* Schools Council Bulletin 23, London: Evans/Methuen, 1971, p 12

[4] *Examinations at 16-plus: proposals for the future*, Joint Examination Sub-Committee of the Schools Council, London: Evans Bros., 1975, p 19

[5] Department of Education and Science, *School Examinations: report of the steering committee established to consider proposals for replacing the GCE O level and CSE examinations by a common system of examining* (Waddell Report), Part I. London: HMSO, Cmd 7281, 1978, para 30

> *(iii) common papers taken by all candidates, but containing questions/part-questions with stated different mark weightings (such as tariff questions) which involve choice of question on the part of the candidate;*
>
> *(iv) a common paper taken by all candidates, plus alternative papers, reflecting different approaches to the subject and/or different forms of assessment, but which are not intended to be at varying levels of difficulty. Candidates can attain the highest grades whichever papers they choose;*
>
> *(v) a common paper taken by all candidates, plus alternative papers which are intended to be at varying levels of difficulty. If the candidate chooses an easier alternative paper he cannot normally attain the highest grades*.[6]

As the Report acknowledged, for the candidates and their teachers there was an important distinction between the approaches to examining described at (i) - (iii) and (iv) - (v) respectively.

> *The former categories (common examinations) do not require the candidate to choose between alternative papers, although they may involve him in choosing between questions at various levels of difficulty or carrying different marks. This kind of choice is exercised by the candidate on the day of the examination usually with previous guidance from the teacher. On the other hand, categories (iv) - (v) require the candidate and teacher to choose between papers, perhaps early in the course of preparing for the examination.* [7]

One of the particular difficulties with alternative (v) noted by the steering committee was that of comparing performance on differentiated papers intended in one case to consist of difficult questions and, in the other, of easy ones. The difficulty was exacerbated when, for example, a poor performance on the hard paper had to be judged against a good performance on the easier one. To address this, the Report commended the practice adopted in some of the joint examinations of having some papers, or other elements of assessment, common to all candidates. It also addressed the issue of the need to express the results for all candidates on a single grading scale.

> *All in all, we are satisfied that the techniques exist*[8] *to overcome the difficulties of assessing candidates who have taken differentiated examinations at different levels of difficulty, and of expressing the results on a common grading scale. Whatever the approach used, the most important ingredient in comparing performance is likely to be the judgement of examiners.*

The reference quoted in the Waddell Report was to the work of Backhouse. Shortly after the publication of the report, however, Backhouse's methods of scaling were criticized by Wood[9], and subsequently a range of methods were proposed in the literature.
(Section 3 of this report reviews the relevant literature)

[6] Waddell Report, Part II, para 20
[7] Waddell Report, Part II, para 21
[8] Discussed by, for example, the Report noted in Backhouse, J K (1976): *Determination of Grades for Two Groups Sharing a Common Paper* (Educational Research Vol. 18 No. 2).
[9] Wood, R. (1978). 'Placing candidates who take different papers on the same mark scale,' Educational Research, 20, 210-15.

The notion of differentiated papers first appeared in government policy in the 1978 White Paper:

> *The educational studies carried out by the Steering Committee led them to the conclusion that in at least some subjects it would be necessary to provide a variety of alternative examination papers and tests, at different levels of difficulty, in order to provide satisfactorily for candidates from the intended wide ability range. This is especially the case where, as in mathematics or modern languages, the range of skills involved is wide or certain concepts are within the grasp of some candidates but beyond the reach of others. The government accept this view, and consider it essential that the examination system should enable all candidates to demonstrate their capabilities. The assessment procedures must, therefore, provide for the inclusion of items suitable only for some candidates, or required only for some candidates, and in such a way that the curriculum is not distorted for others.*[10]

Four years later, the Cockcroft Report on the teaching of mathematics endorsed this view. In chapter 10 of their report, the committee outlined a possible approach to developing an examination in mathematics at 16-plus; but with the caveat that this was only intended to cover the range of pupils for whom the O Level and CSE examinations were designed, i.e. about the top 60 to 65 per cent of the ability range in any subject.

> *The scheme of examination which we wish to outline would provide a range of papers which would enable each candidate to attempt a combination of these papers 'focused' at one of three grades on the scale. For instance, the combination of papers focused at grade 6 would be one on which a candidate who was awarded grade 6 would be able to obtain about two thirds of the marks which were available. A rather higher mark would achieve grade 5 and a lower mark (but not, we would hope, below about 50 per cent) would merit grade 7. It would be possible for a candidate who did exceptionally well and who achieved a very high mark on these papers to be awarded grade 4.*
>
> *We believe that it would be appropriate to provide combinations of papers which would focus at grade 2 (appropriate for candidates expected to gain grades 1,2 or 3), at grade 4 (for candidates expected to gain grades 3, 4 or 5, though with the possibility of achieving grade 2) and at grade 6. We would envisage that the syllabus for the papers focused at grade 4 would be more extensive than that for the papers focused at grade 6 and that some of the questions asked would be of a more demanding type than those included in the papers focused at grade 6. There would similarly be a further increase in syllabus content and difficulty of question in the papers focused at grade 2.*
>
> *It would also be necessary to make provision for the award of grades below those suggested as appropriate for the combinations of papers focused at grade 2 and grade 4. We would, however, expect teachers to advise pupils and their parents as to the combination of papers to be attempted in such a way that only in the most exceptional circumstances would it be*

[10] Department of Education and Science, *Secondary School Examinations: a single system at 16-plus*, London: HMSO, Cmnd 7368

necessary to award a grade lower than 4 on the papers focused at grade 2, and lower than 6 on the papers focused at grade 4...

We believe that it would be possible to implement the suggestion which we are making in a number of different ways. One way would be to set a series of four papers graded in difficulty and content so that, for example, papers 1 and 2 could provide an examination focused at grade 2, papers 2 and 3 an examination focused at grade 4 and papers 3 and 4 an examination focused at grade 6. An alternative arrangement would be to provide three pairs of papers, one pair focused at grade 2, one at grade 4 and one at grade 6, with some questions included in more than one pair of papers if this was felt to be necessary in order to establish comparability between the same grade awarded on different pairs of papers. If our suggestion for the introduction of 'Extra Mathematics' were to be adopted, the relevant paper would be available only to candidates attempting the papers focused at grade 2.

We wish to point out that the provision of a paper which would be taken by all candidates would not accord with the fundamental principles which we have set out [that the examination papers and other methods of assessment which are used should (i) be such that they enable candidates to demonstrate what they do know rather than what they do not know; and (ii) should not undermine the confidence of those who attempt them] *unless it was suitable for inclusion in a combination of papers which was focused at grade 6.*[11]

In putting them forward, the committee recognised that their proposals might not meet with universal support.

We have been given to understand that there are some teachers who are expecting that the introduction of a single system of examination at 16+ will remove the necessity of advising pupils and parents as to the papers within the examination which pupils should attempt. However, the whole of our argument for a differentiated curriculum implies that the same set of examination papers in mathematics cannot be suitable for all pupils. If follows that those who teach mathematics must accept responsibility for giving such advice. We believe that the scheme which we are proposing will help teachers in this respect by giving them ample scope for formulating appropriate advice and will not require decisions to be taken at too early a stage.[12]

Nevertheless, by the early 1980s the notion that differentiated papers – i.e. alternative papers of varying difficulty for different ability groups – should generally be the norm in the new system had become more or less standard thinking. At some point, however, what was to be understood by 'differentiated papers' seems to have hardened into 'fully differentiated papers', something that had not been one of the five alternative strategies identified in the Waddell report. Somewhere the common element in Waddell's fifth alternative ('*a common paper taken by all candidates, plus alternative papers which are intended to be at varying levels of difficulty*') appears to have largely disappeared. Accordingly, when the then Secretary of State, Sir Keith Joseph, announced the

[11] *Mathematics Counts: report of the Committee of Inquiry into the teaching of mathematics in schools under the chairmanship of Dr W H Cockcroft*, London: HMSO, 1982, paras 524-526 and 528-529
[12] Ibid para 527

introduction of the new GCSE examination in the House of Commons in June 1984 there was, he said: '*a need for differentiated papers or questions in every subject, so that each subject may be taught and examined in a way that reflects the widely differing abilities of candidates more effectively.*'[13]

Whereas views on differentiation had arguably grown less flexible by 1984, views on the ability range of the pupils for whom the new examination was intended appeared more flexible. In 1971 the Schools Council had argued strongly in favour of common examination system covering only the 40-100 percentile of the ability range: *'... it is firmly recommended that in the first instance the common system of examining should be designed to assess performance, subject by subject, from the 40th percentile to the top of the range of ability'.* [14] Wider percentile ranges, for example 5-100 or 5 to something below 100, were considered and rejected: *'on the grounds that it would require the exploration of a considerable new area of assessment in which experience is at present very limited, and in which some largely novel methods of assessment might well need to be devised'.*[15]

Over a decade on, however, government policy appeared to take a rather broader view.

> *The examination* ***would be open to all*** (our emphasis) *who wish to take it but the syllabuses and assessment procedures would be designed in order to reward the attainment of pupils whose marks would be likely to place them in the top 60 per cent of candidates if an examination in that subject were attempted by pupils spanning the whole range of ability in the school population.*[16]

This concept of an examination with unrestricted entry but constrained exit points was re-affirmed at the time GCSE was introduced. Asked whether the new system would cover the same ability range as the old CSE and O-level examinations had been designed for – i.e. 60 per cent of the relevant age group - the Secretary of State replied that:

> *The proportion of the population for whom the new examination system is proposed is 100 per cent. The proportion of people who will achieve graded results within it depends ... on how they perform.*[17]

And, in his accompanying letter to Parliament, Sir Keith Joseph made it clear that whilst entry might be open to 100 per cent of the population:

> *... the standards required of successful candidates in GCSE examinations should be no less exacting than those required in the existing 16+ examinations which, taken together, were originally designed for the upper 60% of the ability range.*

[13] Hansard, HC Deb 20 June 1984 vol 62 cc303-13
[14] A common system of examining at 16+, Schools Council Examinations Bulletin 23, London: Evans Bros. and Methuen Educational Ltd, 1971, p 10
[15] Ibid, p 9
[16] Department of Education and Science & Welsh Office, Examinations at 16-plus: A statement of policy, HMSO, November 1982
[17] Hansard, HC Deb 20 June 1984 vol 62 cc303-13

When they emerged early in 1985, however, the General Criteria for the GCSE signalled a move away from the insistence on the 60 per cent figure that had been the target of GCE O-level and CSE examinations. GCSE was not to be similarly limited – although the standards required would be no less exacting.

> *It will be designed, not for any particular proportion of the ability range, but for all candidates whatever their ability relative to other candidates, who are able to reach the standards required for the award of particular grades.*[18]

These required standards would be defined by grade criteria which were then in the process of being developed for each subject.

But, despite the claim that GCSE would not be limited by percentages, the requirement that standards should be no less exacting than previously, meant in practice that the new subject grade criteria would still have to be designed in such a way as to exclude the bottom 30 – 40 per cent of the ability range (although not necessarily the same 30-40 per cent in each subject). Uncontrolled entry to the examination in a given subject could, therefore, result in a high proportion of candidates emerging with nothing to show for the experience. To avoid this, schools would need to continue to act as a filter, selecting for entry only those pupils likely to meet the subject grading requirements. This was so whether or not the examination for which they were entered included differentiated papers.

Many minor changes apart, the GCSE format has remained basically the same since its inception. Initially, most examinations had two tiers: Higher, offering grades A-E (A*[19] from 1994); and a lower (Foundation) tier offering Grades F-G. In 1998, the Higher tier was modified to cover grades A*-D, with the Foundation tier covering grades C-G in addition an 'allowed' Grade E was introduced to the Higher tier for candidates narrowly missing a Grade D.

For many years, mathematics was an exception with three tiers: Higher (grades A*–C), Intermediate (grades B–E) and Foundation (grades D–G). However, mathematics moved to the standard two tier system in 2006 (for the first examination in 2007 or 2008 depending on whether the modular or linear course was taken).

3. Grading differentiated examinations

Where differentiated papers have been used in GCSE and other examinations this has brought in the specific challenge of grading candidates fairly on a common grading scale when their marks have been achieved on either harder or easier differentiated papers. There is no simple solution to that challenge, but in this section, we review some of the approaches that have been tried, going back over a period of more than 35 years.

[18] Department of Education and Science, *GCSE: The National Criteria*, London: HMSO, 1985, para 9
[19] The A* grade was introduced 1994. The threshold for achieving an A* has varied considerably over time, coming down as low as 47% in a 2005 AQA Business Studies GCSE.

(i) Backhouse (1976)

Backhouse's investigation used raw test scores from two groups of candidates[20]: a CSE group (who had sat papers designated **paper one** and **paper two**) and a GCE group (who had sat **paper two** and **paper three**). Backhouse therefore aimed to compare methods of determining grades for two groups sharing a common paper, and to investigate to what extent candidates might be awarded different grades if different methods were used. Backhouse described three methods: ranking, scaling and regression. These are described below.

Ranking

The ranking method described by Backhouse is based on the following assumption: that candidates are best ranked according to their total scores on the two papers offered. The method follows three stages, as follows:

Stage One

- The total score for each candidate is calculated by summing their scores on each of the two papers they attempted. Candidates in the two groups (CSE and GCE) are then ranked separately based on this total score. So now the CSE group are in rank order and the GCE group are in rank order. These rank orders are noted.

Stage Two

- Now **all** the candidates are ranked using their scores on paper two (the common paper). This gives a single rank of all candidates based purely on paper two.
- Now the 'paper two ranks' calculated for each candidate are allocated to the group to which that candidate belongs (i.e. either CSE group or GCE group). This gives a set of rankings for each group, based on candidates' scores on paper 2. For example, exemplar rankings for nine candidates (four in the GCE group, five in the CSE group) may be:
    - CSE: 3,4,5,7,8
    - GCE: 1,2,6,9

Stage Three

- The final rank order is based on combining the rank calculated in stage two (based on all candidates' scores on the common paper) with the ranks calculated within each group (based on candidates' total scores across both papers attempted). This is best explained by example. Using the illustrative ranking data shown in stage two above:
    - Based on the ranking of candidates carried out in stage two, the rank order 1 candidate should come from the GCE group. Therefore the **1st ranked candidate from within the GCE group** is selected as the first ranked candidate overall. Note that this is the candidate who was ranked first in stage one (based on their total score across both papers attempted), and therefore **may or may not be the candidate who was ranked first according to their paper two score**; the paper two score has been used to determine how the candidates from the two groups should be interleaved with each other, but in creating the final rank this method ensures

[20] The data was provided by the East Midland Regional Examinations Board.

that the group rankings calculated in stage one (based on total score over two papers) are maintained.

- This process is repeated for each rank position in turn (i.e. 2 – 9 with the illustrative data).

Scaling

The scaling method described by Backhouse is based on one proposed by Peaker in a private communication to Backhouse. It is based on equalizing the means and standard deviations of scores for each population on each paper. So in Backhouse's study:

- The scores of the CSE group on papers one and two are scaled to have equal mean and standard deviation.
- The scores of the GCE group on papers two and three are scaled to have equal mean and standard deviation.

The CSE group and the GCE group will most likely now have different means and standard deviations from each other, but this is to be expected given the different abilities of the two groups.

Once the marks are so scaled, then a total score for each candidate is calculated by summing their (scaled) score on each paper they sat. Total scores are then used in ranking (and subsequently grading).

Regression

The regression method described by Backhouse involves using the scores achieved by each candidate on the two papers sat (one common and one not) to estimate the score s/he would have achieved on the third paper. The estimate is achieved by deriving an equation to predict what a candidate might have scored had s/he sat for a paper. For example, with the two groups in Backhouse's study:

- The CSE group's scores on papers one and two are used to derive an equation[21] which estimates a score on paper one from a given score on paper two.
- This equation is then used to estimate a score on paper one for each candidate in the GCE group based on their score on paper two.
- The same process is repeated to estimate a score on paper three for each candidate in the CSE group.

In this way, each candidate will have a score for each paper: two actual and one estimated. All three scores are summed to give a total mark for each candidate, which is used in ranking and grading.

Comparison and assessment of methods

Using data[22] provided by the East Midland Regional Examinations Board, Backhouse calculated the grade for each candidate using each of the three methods (ranking, scaling, and regression). He then

[21] Backhouse used linear regression – i.e. the equation derived assumed a linear relationship between the scores on paper X and paper Y.
[22] 157 candidates in the CSE group; 234 in the GCE group; 391 in total.

compared the grades awarded by each method to the actual grades awarded by the Board, and investigated differences. He concluded that the agreement between the Board grades and the ranking method was best (54 total disagreements, which equates to 13.8%). In two cases there was a disagreement of two grades. There were 60 disagreements (15.3%) between the Board grades and the scaling method (three disagreements of two grades) and 61 disagreements (15.6%) between the Board grades and the regression method (one disagreement of two grades, one disagreement of three grades). Where there was a serious discrepancy (i.e. two or more grades) between the Board grade and the grade awarded by one of the three statistical methods, this was often seen to be where the Board would appear to have used professional judgement, and/or have weighted scores differently on the differentiated papers. An example is quoted where a GCE candidate scored poorly on the common paper (paper two) but very well on paper three. The C grade awarded by the ranking method would appear justifiable based on the raw statistics and the importance of the common paper score for the statistical methods. However the Board awarded an A, possibly because the candidate's aggregate score took him/her above the grade A boundary, or possibly because scores on paper three were more highly weighted by the Board.

Backhouse concludes that the ranking and scaling methods have the best theoretical justification, and that it should be a 'matter of convenience' which method is to be used.

There is one further point which worth noting at this stage: **there is no objective metric which can be used to assess the efficacy of methods used to determine grades for two groups sharing a common paper**. That is, when a particular method is applied and a rank order and/or grade calculated for each candidate, there is no 'correct' answer to compare against. Backhouse compared each of the three methods to each other and to the grades awarded by the Boards, but there is no objective way to say than any of these methods is more accurate than any other. Discrepancies between the outcomes of the methods may be investigated individually, looking at test scores, but any attempt at determining which is 'correct' ultimately relies on subjective judgement. This applies not just to the methods described by Backhouse, but to all the techniques which have been developed to place candidates who take differentiated papers on a common scale.

(ii) Wood (1978)

As noted previously, shortly after the publication of the Waddell Report, Backhouse's methods of scaling were criticized by Wood[23]. Wood suggested that it is necessary to ask question like 'Do these methods give placements which accord with common sense?', and 'Are particular pairs of candidates ordered 'the right way round'?' Wood highlights examples of candidate ordering from within Backhouse's study which, in his view, do not accord with common sense or do not order pairs of candidates the 'right way round'.

Wood's criticisms may be summarised as follows:

- For the ranking method, a candidate's final rank depends on how others in his group perform on the common paper or papers, which makes the process rather volatile; small changes in a candidates marks can lead to seemingly large changes in rank order not only for the candidate, but for other candidates as well.

[23] Wood, R. (1978). 'Placing candidates who take different papers on the same mark scale,' Educational Research, 20, 210-15.

- For the scaling method, the criticism is that the distribution of marks between groups before scaling may be quite different, and that if one groups is (for example) skewed and the other is not, then a candidate's final rank will be affected by the nature of the distribution of marks of his group.
- The regression method is based on what Wood calls 'imputed' scores (i.e. one score for each candidate is calculated from their scores on the other two papers), which is unlikely to be acceptable to stakeholders were it to come under public scrutiny.

Wood goes on to suggest that a method of **paired comparisons** is superior. Paired comparisons, as described by Wood and Wilson[24], regard candidates as contestants in a competition who compete against each other as often as they are assessed by the same instruments of comparison. For example, using the same scenario as described above by Backhouse, all candidates are compared against each other on paper two (the common paper), CSE candidates are compared against each other on paper one, and GCE candidates on paper three. The technique is question-mark-based rather than paper-mark-based. It proceeds by comparing the number of times each candidate is better than each of his peers on each question and so gradually refines the rank order and the value to be attached to each measure of the candidates' abilities. A final rank order is arrived at by applying a procedure which credits individuals according to the number of times they gain preference over other candidates, where allowance is made for the quality of those over whom the candidate is preferred (the procedure is complex, and is described in more detail in Wood and Wilson (1978).[24]

Although Wood's notes in the conclusions that 'no one method will give perfectly sensible results in all circumstances', it is worth emphasizing again here that there is no objective metric which can be used to assess the efficacy of methods used to determine grades for two groups sharing a common paper, and that the criteria Wood uses to criticise Backhouse's methods are highly subjective.

(iii) Kingdon (1983)

Wood's suggested method of paired comparisons was itself criticised by Kingdon *et al*[25] which reported a study which attempted to replicate Backhouse's work using data from a larger sample of candidates sitting an operational examination and to implement the alternative technique outlined by Wood.

Kingdon outlines conceptual difficulties associated with techniques based on rank orders (as proposed by both Backhouse and Wood), referencing a result by Arrow (1951)[26] which states that:

- any method based on rank orders has the potential to be unfair to some candidates;
- it is theoretically impossible to recognize situations where unfairness may arise *a priori*.

[24] Wood, R. and Wilson, D.T. (1978). 'Determining a rank order when not all individuals are assessed on the same basis'. In: VAN DER KAMP, L.J. Th. et al. (Eds) Psychometrics for Education Debates. London: John Wiley.
[25] Kingdon, J.M. et al. 1983 Awarding grades on differentiated papers in school examinations at 16+, Educational Research, 25, 3, 220-229.
[26] Arrow, K. J. (1951). Social Choice and Individual Values. New York: John Wiley.

Kingdon also critiques the paired comparisons method proposed by Wood. The technique only uses the fact that A is better than B and fails to consider by how much: considerable loss of information may therefore be involved. Kingdon also questions whether the technique is guaranteed to converge to a final result – at the time of writing no large scale test of paired comparisons for examination data had been carried out[27]. Kingdon also notes the weight attached to candidates' performance on the common paper in the methods proposed by Backhouse and Wood, and also highlights the adverse effect that a common paper which favoured one group over another (due to, for example, question type or content) could have on the grades awarded.

Kingdon's study, using data provided by the Welsh Joint Education Committee from a large sample of candidates sitting an operational examination, produced a number of interesting outcomes:

- Differences between the statistical methods and the actual award made by examiners suggest that the examiners attached less weight to the Paper 2 (common paper) attainment than the purely statistical procedures do.
- Although none of the three methods suggested that any candidate should be assigned a grade which was more than one grade different from the grade actually awarded, there was considerable disagreement between the methods, and the regression method was deemed the 'most idiosyncratic' in its suggested adjustments to grades.

It should be emphasized that Kingdon's assessment of the three statistical techniques proposed by Backhouse is based on a comparison with the judgments of the WJEC chief examiners. Kingdon justifies this as 'judgments appear to have been acceptable to the candidates and their teachers for a number of years'. As before, there is no genuinely objective metric which could have been applied; hence the comparison with existing practice.

Kingdon concludes that statistical methods should not in fact be relied upon to determine grades for groups taking differentiated examinations, but that they could be used to provide additional data for examiners to draw upon when making their judgements.

(iv) Good and Cresswell (1988a)

Good and Cresswell[28] report the results of a study which compared the grades obtained using two scaling methods against those obtained by grading the versions of the examination separately. At the time the paper was written (1988), separate grading of differentiated papers had reportedly

[27] Of related interest here is a paper by Pollitt and Crisp (*Alastair Pollitt and Victoria Crisp, 2004, Could Comparative Judgements Of Script Quality Replace Traditional Marking And Improve The Validity Of Exam Questions?*) which investigated the use of Thurstone paired comparison judgements to avoid the need for traditional marking, with rank ordering of scripts and grading done at the same time. In this case teachers' rank ordering of students 'quality' was used as a criterion measure to assess the efficacy of the paired comparison method and a school's mock exam which was marked traditionally. The correlation between the teacher's rankings and the paired comparisons method was found to be stronger than the correlation between the teacher's rankings and the mock examination, suggesting (assuming that the teacher's rankings were accurate) that the paired comparisons method provided more valid results than the traditionally marked examination.

[28] Good, F.J. & Cresswell, M.J. 1988 Placing candidates who take differentiated papers on a common grade scale, Educational Research, 30, 3, pp177-189

been generally adopted because (according to Good and Cresswell) 'it is a method that is much closer to that used in the O-level/CSE dual system'.

The scaling methods, by comparison, place candidates who take differentiated papers on a common grade scale. Grade boundaries are set on the constituent papers as before and then scaled onto the common grade scale in the same way as candidate's raw scores. Grade boundaries on the common grade scale are then aggregated by simple arithmetic (i.e. they are averaged).

The two scaling methods used were equipercentile scaling and linear scaling. Linear scaling is essentially the same technique described by Backhouse. Equipercentile scaling also scales scores on an alternate paper to the mark scale used on the common paper, but makes no assumption about the relationship between the achievements measured by the papers (linear scaling assumes a linear relationship).

The data in the study came from the Novel Examinations at 16+ Research Project[29]. Data from two 'experimental examinations', one in history and the other in physics, were used. In history there were 173 General Level candidates and 104 Extended Level candidates; in physics there were 120 candidates entered at the General Level and 84 at the Extended Level.

As in earlier studies, Good and Cresswell compared the statistical techniques to each other and to the grades awarded by examiners. However, they also devised an 'external measure' against which to consider all three approaches: a grade was derived for each candidate using prior attainment data. For both subjects in the study (history and physics) the correlations of the scaling methods and the external measure were higher than the correlation of the separate grading technique (i.e. the one actually used for grading by examiners) and the external measure.

Good and Cresswell conclude that:

- Grading using the scaling techniques was at least as good and possibly slightly superior to separate grading in terms of the relationship to the achievement measured.
- Significantly, adopting a scaling method was seen to address a reported issue with separate grading which suggested that grades available from more than one version of the examination may be more readily attained from the easier version(s) of the examination (see Good and Cresswell 1988b below). This became known as the **Good and Cresswell effect**, and has been recognised as a major issue in terms of having confidence in any judgemental process for grading differentiated papers onto a common scale.
- The authors also state their belief that scaling is to be preferred as the final grade scale is likely to be 'more satisfactory' (in a technical sense) if one mark scale is partitioned into one grade scale than if points on two separate mark scales are linked in order to arrive at one grade scale. Concrete examples of this include:
    - On a single grade scale components are given their intended weights and, equivalently, the scaled marks represent equal increments of achievement on both components.

[29] SEC (1986). Novel Examinations at 16+. London: Secondary Examinations Council.

- On a single grade scale each grade has only a single 'width' (range of marks), which is not necessarily the case when two separate mark scales are linked.
- From a technical point of view, grade limitations are unnecessary after scaling, which could address the issue of candidates who are inappropriately entered for the more difficult versions of differentiated papers examinations, or where candidates are inappropriately entered for the less difficult version and their grade is capped by no matter how well they score.

- The differences in performance between equipercentile scaling and linear scaling using the available data were not significant, but linear scaling was the recommended approach as equipercentile scaling has the potential to compress extreme marks.

(v) Good and Cresswell (1988b)

In a second 1988 paper[30] Good and Cresswell raised doubts about whether the processes by which examiners were determining grade boundaries at the time were appropriate for differentiated papers examinations. Specifically they considered whether teams of awarders make grade awarding judgements that are consistent, and whether judgementally determined comparable standards and statistically determined comparable standards are compatible. The statistical definition of comparability used is that, between components, equal proportions of the same group of candidates should lie above each boundary[31]. The authors quote data which indicates that where decisions are made based on purely professional judgement, a higher proportion of candidates tend to reach each grade boundary on easy components than on harder ones. This, as noted above, became known as the Good and Cresswell effect, and has been recognised as a significant result in differentiated examinations.

The authors conclude that the only way to ensure that candidates, regardless of level of entry, stand an equal chance of reaching each of the grades on differentiated papers is to adopt a purely statistical definition of comparability which requires the same proportion of candidates to reach the grade boundaries on each of the components. The authors note that this assumes that the common paper is an equally valid measuring instrument for all candidates, and that this may not always be the case. They conclude that in practice, therefore, it may be desirable to use composite procedures to reach comparability, with the awarders fixing the boundaries judgementally but in full knowledge of the effects this will have on the statistical comparability of the grades. This would seem to be the model which was ultimately adopted by awarding organisations for GCSE.

(vi) Stobart *et al* (2005)

Stobart *et al*[32] report the results of an independent evaluation of two schemes of 'two-tier' GCSE mathematics (the 2005 GCSE Pilot and Trial in mathematics respectively) which were examined

[30] Good, F. J. and Cresswell, M. J. 'Grade awarding judgements in differentiated examinations' British Educational Research Journal, 14, 261-279

[31] For example, referring back to Backhouse's study, this would mean that equal proportions of the CSE group should lie above each boundary, and that equal proportions of the GCSE group should lie above each boundary, but not that the proportions are the same at each boundary for each group.

[32] Gordon Stobart, Tamara Bibby and Harvey Goldstein, 2005, Moving to two-tier GCSE mathematics examinations: an independent evaluation of the 2005 GCSE Pilot and Trial.

alongside the three-tier model which was in operation at the time. Of interest here is how the grades were arrived at in the Pilot and the Trial.

In the **Pilot**, candidates took two adjacent papers, each covering a limited range of grades (A*-B; C-D; E-G, see Figure 1) from which the best grade was selected. There was no overlap between papers, and the marks from the two papers were not combined; the information from one paper was ignored. The grade was the best result from the two separate papers[33].

In the **Trial**, candidates were entered for either the Foundation or Higher tier. Papers covered overlapping grades (A*- D and C-G, see Figure 1). Raw marks from both papers were converted onto a uniform mark scale (UMS), as were the coursework marks, and combined to generate the final grade. The awarding process for the Trial, as described by Stobart, was based on what was referred to by Good and Cresswell as 'separate grading' – grade boundaries were set independently for each paper. The actual boundaries themselves do seem to have been based on a statistical definition of comparability; the percentage of candidates at grade C on each paper was kept broadly equivalent, which is broadly in line with that recommended by Good and Cresswell. Statistical information was also used, in conjunction with examiner judgements, to set the A boundary, and the A* boundary was statistically determined to equate the percentage gaining A* with that of the other GCSE taken (candidates on the Trial sat both the three-tier and the Trial examinations with the best final grade being certificated).

The conversion of marks to UMS in the Trial has both similarities to and differences from the scaling proposed by Backhouse and subsequently recommended by Good and Cresswell. In converting to UMS, grade boundaries and raw scores from component examinations are transformed onto a common scale (using a non-linear transformation). However the scaling involved is determined by a) the grade boundaries on the component examinations and b) the predetermined GCSE grade boundaries agreed by inter-awarding body agreement[34]. There is no concept of using scores on the common paper to calibrate the performance of the two groups: in the UMS transformation, each paper is assumed to be independent.

[33] Note that failure to get a D on the middle paper meant that the grade from that paper was a U, so that any legitimate grade from the E-G paper was better than that.

[34] By inter-awarding body agreement, the uniform mark grade boundaries in GCSEs are at the following percentages of the maximum uniform mark for the unit/module or qualification: A* 90%, A 80%, B 70%, C 60%, D 50%, E 40%, F 30%, G 20%

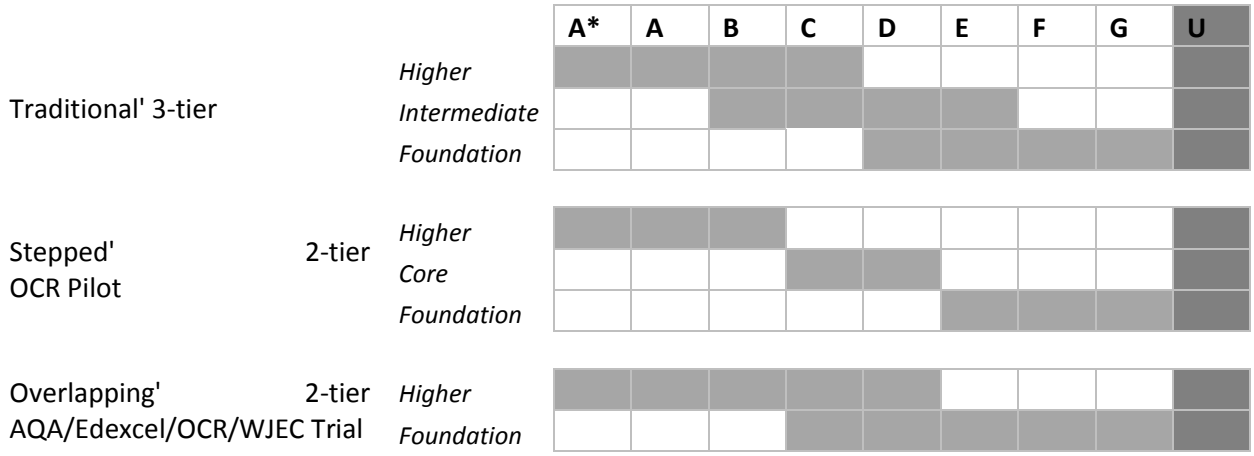

Figure 1: Models of tiered GCSE mathematics examinations evaluated by Stobart et al, 2005

(vii) Coe (2008)

Coe reports a study of the comparability of examinations in different subjects which attempts to place grades from different GCSE subjects onto a common scale to estimate their respective difficulty. Coe used the Rasch model to analyse data from over 600,000 GCSE candidates from 2004. Of the 109 GCSE subjects in the full dataset, 34 were found to fit[35] the Rasch model (i.e. 75 did not), and so were deemed to be measuring a common construct. Coe argues that this might be considered to be a trait such as general academic ability[36], rather than 'difficulty' – the concept of comparing the difficulty of different subjects having received substantive criticism in the literature.

The Rasch model would seem to have some potential for placing candidates who take differentiated papers on the same mark scale, and potential models are discussed by Wheadon[37]. However as Coe points out, the use of Rasch has been controversial in the UK. Wheadon further points out that a suitable technical infrastructure is required to support such analyses, as well as appropriate levels of psychometric expertise.

Conclusion

If EBCs do not have tiered papers then the methods discussed above will not be needed. However, given our doubts over whether non-tiered EBCs in all the core subjects is a realistic ambition, we have included this section of the review to inform discussions about comparable grading across tiered papers, as these might arise. There is no one single best approach and some of the choices outlined here will interact with the parallel discussions which are taking place about which approaches will be used to standard set and grade EBC examinations, whether or not they contain tiered papers.

[35] Goodness of fit being determined using a threshold of 1.7 for the infit and outfit residuals.
[36] It is debatable whether the concept of "general academic ability" adds anything to the fairly general point that people who do well in one subject tend to do well in others (even across those subjects most often seen to depend partly on aptitude, such as mathematics and languages).
[37] Wheadon, Christopher Brian (2011) An Item Response Theory Approach to the Maintenance of. Standards in Public Examinations in England, Durham thesis. http://etheses.dur.ac.uk/615/1/Chris_Wheadon_PhD.pdf

4. Non-tiered examinations at GCSE: the case of history

Although common to most subjects, tiered papers have never been used in GCSE history - despite occasional calls for their introduction.[38] The general criteria for the new GCSE required that 'all examinations must be designed in such a way as to ensure proper discrimination so that candidates across the ability range are given opportunities to demonstrate their knowledge, abilities and achievements: that is, to show what they know, understand and can do'.[39] For history the Awarding Organisations opted to meet these criteria by setting common question papers to be answered by candidates of all abilities and awarding grades on the basis of candidates' responses – i.e. 'differentiation by outcome'.

In some quarters, however, there was some hesitancy about adopting such an approach. There were perceived difficulties in seeing how it could ensure that all grades would be rewarded on the basis of positive performances. There were fears that it would perpetuate the drawbacks of past common-paper examinations with the less able obtaining low grades based on few marks; the most able not having opportunities to demonstrate all that they knew and could do; and the average being awarded grades in the middle of the range gained from on a random selection of around 50 per cent of the marks available.

(i) Secondary Examinations Council (1988)

Given reservations of this kind, most notably about the proposals for history and English, the Secondary Examinations Council (SEC) investigated the claims made for differentiation by outcome. During the latter half of 1985 two small committees, one for history and one for English, assembled evidence presented by the examining groups and reported to SEC in October 1985.[40] The reports argued that differentiation by outcome was a viable technique – provided that the questions and the marking practices met certain requirements. Further, in the view of the committees:

- Differentiation posed considerable problems, whatever technique was used. They highlighted in particular issues of establishing hierarchies of difficulty in examinations based on stepped questions and/or differentiated papers.
- Differentiation by outcome, with questions accessible to all candidates, was more compatible with the principle of differentiating on the basis of positive achievement. Differentiated papers, for example, by their nature excluded some and could thus prevent candidates from demonstrating their full potential.

[38] See, for example: 'We therefore recommend that assessment at GCSE should be through tiered examination papers and through coursework.' *Proposals for better history in schools* (undated but c.2010-11) www.anglia.ac.uk/.../betterhistory.../better_history_proposal.pdf

[39] Department of Education and Science, *GCSE: The National Criteria*, London: HMSO, 1985

[40] Secondary Examinations Council, *Differentiation by outcome in history: an abridged version of a feasibility study and report to the Secondary Examinations Council*, October 1985, mimeo, London: SEC; and *Report of the English working party on the differentiated assessment in common papers project*, October 1985, mimeo, London: SEC

Following these reports, SEC established a working party to investigate further the issue of differentiation by outcome in GCSE history examinations. Their conclusions were published in May 1988.[41] These were not intended to be prescriptive, but rather to raise the awareness of history examiners. Additionally, the Working Party felt that their report might be of use and interest to examiners in subjects other than history – particularly those that were also aiming to differentiate in a similar way.

In the report the Working Party noted that some history examinations, notably the national inter-board O level/CSE Schools History Project (SHP) examinations, had been developing a distinctive approach to differentiation by outcome over a period of some ten years. Initially, the rationale behind this approach had not been the issue of differentiation as such. Rather, it had been the perceived need to have a form of questioning and marking that was sufficiently flexible to allow for a range of possible rewardable responses – in relation, for example, to the erection and testing of hypotheses. This recognition of the need for such an approach was not new - nor was it confined to history. It had, for example, been identified some 20 years previously in relation to examining Nuffield Biology: '*If a candidate is asked to propose a hypothesis and thinks of one that is not in the mark scheme, is it to go unrewarded? Obviously not; to do so would make nonsense of the "inquiry" approach.*'[42] With the introduction of joint O level/CSE examinations the requirement became more pressing to have questions which were accessible and meaningful to candidates across a wide range of ability, and mark schemes which were able to categorise the wide range of answers produced.

The main features, identified in the report, of the questions and mark schemes used in SHP based examinations in history were:

i. Questions (and any data on which they were based) that were designed to be accessible to the full GCSE ability range.
ii. Questions that were not reliant exclusively on a candidate's ability to recall one or more specific items of information in order to be attempted.
iii. Questions which admitted of a range of possible responses (i.e. questions at the 'open' rather than the 'closed' end of the spectrum).
iv. Mark schemes which categorised responses in a number of 'levels'. The levels were hierarchical and descriptive of the type of response expected at a given level.[43]
v. The levels anticipated at the question/paper setting stage were later modified on the basis of reviewing candidates' answers in the examination itself. Although there were differences between boards in the comprehensiveness of the reviews that were undertaken and in the number of levels developed for any one set of responses, the principle was that marking guidelines should be essentially 'post hoc' (i.e. be informed by actual rather than anticipated answers).

[41] Differentiation by outcome in history, Report of an SEC/Joint Council Working Party, May 1988
[42] JJ Head, *Flexibility in interpretation of an O level mark scheme*, Educational Review, 19, 2, February 1967
[43] Such mark schemes were the forerunners of the current 'levels of response' schemes used in GCSE and GCE A level history examinations.

In raising issues and presenting the Working Party's conclusions, the report followed the general chronology of developing an examination paper and marking scheme. This was on the grounds that:

> *An attempt has been made to portray a complete process, since the Working Party is of the opinion that it is not a single procedure or feature within the process which achieves differentiation by outcome, but the interaction of all the features described.*

Importantly, the Working Party took the view that the '*adoption of only one or two of ... [the procedures or features] will not ensure differentiation; it will be necessary to follow the complete process ...*'[44]

(ii) Good (1989)

In the context of the Novel Examinations at 16+ Research Project, Frances Good raised a number of concerns about setting common papers that sought to differentiate.[45] Central to the study was the problem, as Good perceived it, which arose in almost all GCSE examinations - with the notable exception of mathematics. This was the requirement for examiners to set some papers which were appropriate for testing candidates from the full examined ability range in that subject.

The study considered two means of differentiating on a common paper: (i) by setting questions based on an incline of difficulty; and (ii) using neutral stimulus questions and seeking to differentiate on the basis of outcome. Although many of her specific reservations were directed towards physics and French and not history, and although all the examinations proved to be adequate discriminators, Good concluded that the first of these methods did not differentiate effectively.

> *It seems clear that this method* [questions based on an incline of difficulty] *of providing differentiation n common papers is ideal for neither of the extreme groups of candidates. It is doubtful whether common papers with a few questions or parts targeted at each level of achievement would genuinely allow all candidates sufficient opportunities to show what they know, understand and can do.*[46]

Further, the study identified a number of practical and epistemological problems associated with setting questions based on a perceived incline of difficulty.

The experimental examinations used only one paper (in history) that sought to differentiate by outcome. Here, given a number of caveats, the conclusions were marginally less pessimistic that effective differentiation could be achieved in this way. Two of the more intractable problems the study suggested, however, were ensuring that:

[44] Differentiation by outcome in history, op cit, p 4
[45] Frances Good (1989): Setting Common Examination Papers that Differentiate, Educational Studies, 15:1, 67-82
[46] Good (1989) op cit, p 72

- candidates respond to questions at the highest level of which they are capable when the questions are designed to be accessible and to be answered appropriately at a number of different levels;
- marking schemes are able to reward a range of possible responses from high to low (usually via banding rather than fixed-mark schemes[47]) without jeopardising inter-marker reliability.[48]

In some ways, the potential problems with attempting to achieve differentiation by outcome that the study encountered reflect the danger that the SEC had warned of: namely that of adopting only one or two aspects of what should be a multi-faceted process.

Whilst concluding that in order to ensure genuine differentiation '*it is almost certainly necessary to use differentiated papers examinations in which candidates only take the papers which are designed to test their level of ability and achievement*',[49] the study did acknowledge the difficulties that these examinations too presented. In addition it recommended the need for more research into the use of differentiation by outcome to establish whether or not it could provide a sound basis for examining across a wide ability range.

5. Where the case for tiered examinations is stronger

Mathematics GCSE examinations are an interesting contrast to the position in history. 'Differentiation by outcome' tends to be the by word in history, and that contrasts markedly with 'differentiation by task' in mathematics. In mathematics GCSEs there has been little debate about whether tiering was necessary. Rather the focus has been much more on how many tiers were required. Stobart, Bibby & Goldstein (2005)[50], for example, provide a detailed account of research which looked at the relative merits of two or three tiers in mathematics. This was around the time when a decision was made to reduce mathematics GCSE from three to two tiers. Looking forward to the EBC, the Advisory Committee on Mathematics Education (ACME) and other mathematics associations have been adamant in their responses to the Consultation Document that tiering is essential to allow students with differing achievement levels in mathematics to be given a reasonable chance to demonstrate what they know, understand and can do.[51] Further, mathematics educators around the world fairly uniformly subscribe to the principle of differentiation by task; and this is reflected in a common approach to setting differing examination papers for students in mathematics depending upon their rate of progress through mathematics curricula.

[47] A 'fixed-mark' marking scheme refers here to a scheme based on allocating a mark to a given unit of information.

[48] RJL Murphy (1982), Further report of investigations into reliability of marking GCE examinations, *British Journal of Psychology*, 52, pp 58-63

[49] Good (1989) op cit, p 81

[50] Stobart, G., Bibby, T., and Goldstein, H. (2005) *Moving to two-tier GCSE mathematics examinations: An independent evaluation of the 2005 GCSE Pilot and Trial (Final report)*. QCA.

[51] *'There is a broad consensus within the mathematics community that having no tiering in mathematics is neither feasible nor desirable in the immediate future'*; ACME response to the Department for Education Consultation on Key Stage 4 Qualification Reform, December 2012, pp 7-8. Similar views were expressed by science educators amongst others. See Appendix.

The logic of the position in mathematics does not seem hard to follow. Good mathematics examination questions tend to clearly delineate mathematical problems, which students are then challenged to complete appropriately. Although a problem exploring, say, a candidate's ability to solve 'differential equations' can lead to a range of responses from different candidates, there is likely to be little merit in the majority of incorrect responses, as there is effectively only one single correct answer to the problem. Giving a single mathematics examination to candidates with wildly differing levels of achievement in mathematics will almost inevitably underwhelm the best candidates, who will romp through it without experiencing much to stretch them; and totally overwhelm the less good candidates, who will end up staring at many of the questions with little or no idea about how to make even a token response to them. The possibility of setting 'differentiation by response' questions seems much more remote and risky. Take, for example, a question like:

> *You have been asked to estimate the volume of water in a local swimming pool, how much it weighs, and how much of it will evaporate over one year. Describe as many approaches that you could use as possible.*

Such a question fails to meet the usual criteria for setting reliable and valid examination questions in mathematics, and will inevitably begin to test skills that strictly speaking are not included in current GCSE mathematics specifications. This situation is as true in Poland, Hong Kong, Malta, Australia and Canada. Mathematics educators are more or less united in their view that the best mathematics examinations are tailored to the achievements of individual pupils. There is also no systematic evidence of tiered examinations on mathematics 'capping students' ambitions', whereas much has been written about the dangers of turning students off mathematics by expecting them to master more advanced mathematical challenges before they have become competent at lower levels of the subject. Indeed 'fear of mathematics' is a well-established phenomenon, which innovative schemes trying to increase success in mathematics have worked hard to overcome.

6. The impact of tiering on students, classrooms and equity issues

Another very important strand of work on the tiering of national examinations has involved studies looking at what happens to individual students within systems where choices need to be made between entering them for different tiered examinations. Clearly this is not a straightforward matter, and there is scope for inappropriate choices to be made in entering students for either too high or too low a tier. In examinations where only restricted grade bands are available for specific tiered papers then this can lead to potentially damaging consequences when students receive say an ungraded result, when entered for the top tier; or a result which does not reflect their actual achievement level, when entered for a lower tier. Such considerations interact with issues of student self-esteem, teacher judgements and stereotypes, and have been shown in several countries to interact with gender effects with girls and boys being treated unequally when it comes to assessing their potential to perform in tiered examinations. A series of reports from Ireland (Elwood & Carlisle, 2003) and Malta (Chetcuti, 201 and 2009)[52], as well as extensive research into the impact of tiering

[52] Chetcuti, D. (2001). Meeting the challenge of equity: the introduction of differentiated examination papers in Malta (pp. 95 – 120). In Sultana, R. G. (Editor). *Challenge and Change in the Euro-Mediterranean Region.*

in GCSE (Elwood & Murphy, 2002)[53] have shown how differentiated examinations of this kind can produce inequalities in the classroom, where teachers may be influenced by factors such as gender stereotypes and effectively incorrectly enter male and female students when faced with a choice of different tiers. A DES (2008) Report on gender effects in education summed up this phenomenon as follows:

> *Elwood (2005)[54] has found that girls and boys are differentially entered for different tiers in Maths (this may be a reflection of prior attainment but may also be influenced by teachers' expectations). She also argues that the tiered entry scheme can have a significant effect on the gender gap.*
>
> *A teacher will decide which tier of a subject a pupil should be entered for, which has implications for the possible range of marks a pupil can achieve. The teacher will make this decision based on prior attainment but it is recognised that such decisions are "a value-laden activity" (Elwood, 2005) ... proportionately more boys than girls are entered for the Foundation (lower) tier in Maths (with a maximum grade D), more girls for the Intermediate tier, and more boys for the Higher tier.*
>
> *Being entered for the Foundation tier is argued to have a different impact on boys than girls (Stobart et al., 1992): Low ability girls are generally better motivated than low ability boys. Boys tend to feel that the lower tier is not worth it. Girls are often more content to take a lower tier. Stobart et al. have also argued that the larger female entry in the intermediate tier represents an underestimation of girls' mathematical abilities by teachers who perceive girls to be less confident and more anxious of failure in maths than boys and more adversely affected by final examinations. The intermediate tier offers grade C while avoiding the risk of being unclassified if performance drops below this grade on the higher tier. Elwood (2005) found that, for GCSE Maths (2003 results), girls achieved 2% more A*–C grades than boys in GCSE maths; however tiering arrangements indicate that more of girls' top grades came from the intermediate tier.*[55]

A recently published review of UK research into a range of issues relevant to the EBC proposals (Baird at al, 2013)[56] includes the following findings, which are all relevant to the debate about tiering:

- French, mathematics and science teachers all are in favour of tiering to stretch the most able students.(Baird et al, 2001)[57]

New York: Peter Lang; and Chetcuti, D. (2009). Identifying a gender-inclusive pedagogy from Maltese teachers' personal practical knowledge. *International Journal of Science Education,* 31(1), 81 – 99.

[53] Elwood, J. & Murphy, P. (2002) Tests, Tiers and Achievement: Gender and Performance at 16 and 14 in England, European Journal of Education, 37(4), 395-416.

[54] Elwood, J. (2005) Gender and Achievement: what have exams got to do with it?, Oxford Review of Education, 31(3) pp. 373-393

[55] (DES, 2008 – Gender & Education: the evidence on pupils in England. DES, London. 2008)

[56] Baird, J., Ahmed, A., Brown, C., Elliott, V. & Hopfenbeck, T. (2013) *World Class Qualification: Research evidence relating to the proposals for reform of the GCSE.* Oxford University Centre for Educational Assessment Report 13/1

- Fifty per cent of teachers in all subjects reported demotivation of students entered for lower tiers. (Baird et al, 2001)

- There is widespread evidence that when teachers set or stream students their decisions are not only influenced by their estimation of ability but are confounded by factors such as socio-economic status, gender and race. (Gillborn & Youdell, 2000)[58]

- There is also evidence that in an attempt to maximise the percentages of students achieving C grades at GCSE, which are critical for league tables, teachers often 'play safe' placing students in a lower tier when they might have had the potential to do better in the higher tier. (Gillborn & Youdell, 2000; Burghes, Roddick & Tapson, 2001;[59] Baird et al, 2001)

This body of research puts a much more complex gloss on the assertion that tiering caps expectations and reduces motivation for those placed in the lower tier. It seems that in reality tiered examinations, which may in many ways boost student engagement and participation in learning and assessment, can face both teachers and students with specific challenges and choices in deciding which tier is the correct one for them to aim at. Like other aspects of teaching and learning, these challenges can be faced in many different ways and certainly do not represent a compelling argument against the use of tiering as a means of providing appropriate assessment opportunities to large national cohorts of students, whose achievement levels may be very difficult to match to a single examination paper.

The problems that are currently perceived as limiting the effectiveness of the current GCSEs will not all be removed by new technical approaches to the setting and grading of EBC examinations. The impact of examinations on students' learning, aspirations and motivation, as well as on classroom practice, teaching strategies - and the public attitudes towards them - depends upon much more than the technical design of the examinations. There is a raft of wider contextual issues - such as school performance tables, government targets and the use of value-added measures to mitigate against crude uninformed interpretations of examination results - which make a huge difference. Consequently, if the EBC is to be a success, that will depend as much on the preparation of a full programme of introducing the new qualification as it will on the technical design of specifications and examination question papers. Such a programme will need to include appropriately resourced teacher preparation sessions, and the provision of information for parents, carers, students, teachers and employers.

7. International Perspectives

In several of the earlier sections of this report we have made reference to work in other countries in terms of the way in which it can contribute to the difficult challenges being faced in relation to the

[57] Baird, J., Fearnley, A., Fowles, D., Jones, B., Morfidi, E. & While, D. (2001) *Tiering in the GCSE.* London, Joint Council for General Qualifications

[58] Gillborn, D. & Youdell, D. (2000) *Rationing education: policy, practice, reform, and equity*. Buckingham, Open University Press

[59] Burghes, D., Roddick, M. & Tapson, F. (2001) 'Tiering at GCSE: Is there a Fairer System?' *Educational Research* 43(2) pp. 175-187

EBC developments. Chetcuti's work in Malta (Chetcuti, 2001 and 2009) is a good example in that it replicates the findings from England about the dangers within a tiered system of teaching and examining of students ending up in the wrong group. This is especially worrying when misplacements of this type appear to be compounded by factors such as gender, race and social class. The challenge of getting young people from disadvantaged backgrounds to achieve their potential within our education system has long been seen as a major priority, and it is worrying if differentiation can create a context within which further barriers are put in the way of such individuals. As we have argued, however, it is unfair to lay the blame for such a phenomenon at the door of those who advocate tiering as the best way to assess large national cohorts of students in certain core subjects.

Another conclusion to be drawn from our international searches is that tiered examinations are fairly common throughout the world, especially in situations where a wide cohort of students is being assessed through a national examination. The distinction we are drawing here is with much more selective examinations such as GCE A-levels taken at a stage in schooling systems where only a high achieving sub-set of students is being examined, principally as a basis for deciding which of them may progress to degree-level studies in higher education.

Differentiated examinations with similar characteristics to GCSE, can be found in Scotland, Malta, Hong Kong, Korea and Finland, for example; and it is notable that the last three on that list all feature as high performing jurisdictions as judged by PISA results. The situation more widely is complicated by the fact that the majority of countries only have formal curriculum based examinations at a later stage for a select group of students who are considered as having the potential to progress to higher education. So, as with A-level examinations, there is far less of a need even to consider tiering because the achievement spread within the group is very much narrower. Another relevant issue is that in a number of other jurisdictions, teacher assessments are the main basis for curriculum focussed assessment, supplemented by much more general university admissions selection tests.

Probably the strongest observation to emerge from the international aspect of our review is that England is somewhat unusual in running elaborate high stakes curriculum-based examinations for both 16 and 18 year olds. The majority of countries restrict the most formal assessments to the point at which students are being considered for entry to higher education, and generally spend a much lower proportion of their resources on tests and examinations.

8. Summary and conclusions

Rationale for different approaches to examining at GCSE: a summary		
Tiered papers	**Non-tiered papers**	**Common paper + alternative papers/questions**
○ Elements of strategy (v) as identified in the Waddell Report (1978), but minus the common paper *(see p 4)* Strong arguments have been put forward in favour of tiered papers in relation to science, foreign languages and especially mathematics from the beginnings of GCSE. ○ The government accepts the view that '*in at least some subjects it would be necessary to provide a variety of alternative examination papers and tests, at different levels of difficulty, in order to provide satisfactorily for candidates from the intended wide ability range. This is especially the case where, as in mathematics or modern languages, the range of skills involved is wide or certain concepts are within the grasp of some candidates but beyond the reach of others.*' (1978 White Paper) ○ *Mathematics is a hierarchical subject ... A concept which some may comprehend in a single lesson may require days or even weeks of work by others, and be inaccessible, at least for the time being, to those who lack understanding of the concepts on which it depends. This means that there are very great differences in attainment between children of*	○ Strategies (i) – (ii) as identified in the Waddell Report (1978) *(see p 4)* Currently, of the proposed EBC subjects, used only in history, although responses from geography and modern languages suggest possibility of some wider support (see Appendix) The main rationale initially put forward in the case of history was that differentiation by outcome, with questions accessible to all candidates, was more compatible with the principle of differentiating on the basis of positive achievement. Differentiated papers, for example, by their nature excluded some and could thus prevent candidates from demonstrating their full potential. The main features of non-tiered papers in history are: ○ Questions (and any data on which they are based) are designed to be accessible to the full GCSE ability range. ○ Questions are not reliant exclusively on a candidate's ability to recall one or more specific items of information in order to be attempted. ○ Questions admit of a range of possible responses (ie questions at the 'open' rather than the 'closed' end of the spectrum). ○ Mark schemes categorise responses in a number of	Strategies (iii) – (v) as identified in the Waddell Report (1978) *(see p 4)* Not an approach currently adopted, but at least some of the responses to the Consultation considered it worth exploring. (See responses by SCORE and ACME in the Appendix)

the same age. (Cockcroft Report, 1982,para 228) ○ *... it will be essential to provide a number of different papers so that candidates may attempt those papers which are appropriate to their level of attainment.* (Cockcroft Report, 1982, para 521) ○ *...there is a very broad range of attainment at Key Stage 4. For these reasons, a single paper for all grades is impracticable, as weaker candidates would be faced with challenging questions they could not begin, and strong candidates would waste time working through questions that present them with no challenge whatsoever* (ACME response to the Consultation on Key Stage 4 Qualifications, 2012, see Appendix)	'levels'. The levels are hierarchical and descriptive of the type of response expected at a given level. ○ The levels anticipated at the question/paper setting stage are later modified on the basis of reviewing candidates' answers in the examination itself. The principle is that marking guidelines should be essentially 'post hoc' (i.e. be informed by actual rather than anticipated answers).	
Advantages/Disadvantages		
(i) Addresses possible motivational issues arising from candidates being faced with tasks beyond their ability and arguably offers a more positive experience. Alternatively, there is concern that entering for lower tiers is itself demotivating. (ii) Consistent with wide range of attainment amongst students of similar age. (iii) Offers appropriate challenge at all levels on the ability spectrum. (iv) The difficulty of comparing performance on papers intended in one case to consist of difficult questions and, in the other, of easy ones. This is exacerbated when, for example, a poor performance on the hard paper has to be	(i) Differentiation by outcome seen to be fairer to students and to be more consistent with the principle of enabling them to demonstrate their full potential. (ii) Bypasses the technical and attitudinal difficulties associated with other approaches. (iii) All the available evidence indicates that non-tiering offers the possibility of a practical, valid and reliable approach only in some subjects.	(i) Might help to address some of the technical and teaching/learning problems associated with fully tiered papers, particularly those associated with comparing candidate performance on papers of different difficulty. The Waddell Report (1978) commended the practice adopted in some of the joint examinations of having some papers, or other elements of assessment, common to all candidates. (ii) Developing a common paper that was appropriately challenging to the full ability range could pose similar difficulties to those associated with fully tiered examinations.

judged against a good performance on the easier one in order to place candidates on a common grading scale. (v) Requires schools to make possibly contentious teaching/learning decisions related to matching students' ability to appropriate paper in advance of the examination.		

8.1 In this paper we have considered in particular the cases of history and mathematics. These two subjects were chosen in order to provide sharply contrasting examples. Other subjects, for example geography, (see Appendix) might lend themselves slightly more readily than mathematics to a 'differentiation by response' approach, but perhaps not as well as history. Science examinations tend to use 'differentiation by task' more than say English literature and geography. So we have a complex picture with differentiation posing different challenges in different areas of the curriculum. From the introduction of GCSE, however, the balance of advantage in achieving effective and fair differentiation in most subjects has been seen to lie in the provision of tiered papers.

8.2 Setting various types of differentiated examinations raises varying problems of validity, pedagogy and comparability of grading outcomes. Given that tiered papers have been the norm, the bulk of the research has focused on issues surrounding the central problem that arises when groups of candidates of different abilities sit different examination papers with questions set at different levels of difficulty, and yet common standards must apply for the award of each grade.

8.3 A number of methods to determine grades for two groups sharing a common paper have been proposed in the literature and reviewed here, including: ranking, scaling, regression and Rasch. It is important to stress again that, although external measures (such as results on other examinations or teacher assessment) may offer one yardstick, there is no objective metric which can be used to assess the efficacy of methods used to determine grades for two groups sharing a common paper; there is not necessarily a 'correct' answer.

It is also worth restating the significance of the result which came to be known as the Good and Cresswell effect (that examiners will find it easier to reward performance on the relatively easier questions posed on the lower-tier examination).

In practice therefore, there is an argument for judicious use of statistical techniques in grading together with human judgement. As Backhouse notes in his comment on Woods' paper, '*If examiners are able to use more than one method, their attention will be drawn to individuals at grade borderlines who are allocated different grades by the two methods. Such candidates clearly ought to have special attention paid to them'*. How practical this is for large cohorts, however, is unclear.

8.4 In the case of non-tiered examinations such as history three criteria appear to be crucial in setting papers if they are to differentiate effectively:

- Questions/tasks must be sufficiently open as to be capable of allowing candidates to respond in a variety of ways that are valid, but which nonetheless are rewarded at different levels within a hierarchical framework.
- Questions/tasks must not use language and constructs in ways that pose unnecessary barriers in the way of their accessibility.
- Candidates must be enabled to respond to questions/tasks at the highest level of which they are capable.

However, all of the evidence that we have reviewed concludes that for many subjects, where there is a wide range of ability amongst candidates, attempting to apply these criteria greatly diminishes the value and meaning of an examination. It has, of course, always been the case that general rules intended to govern good examinations rarely apply to all subjects in the same way.

8.5 We have reviewed both the benefits and the dangers of tiering as a means of providing adequate differentiation in large-scale national examinations. Our overall conclusion is that tiering undoubtedly brings new challenges into the process of fair and equitable grading, and may indeed act against the interests of individual students, especially when they are placed in a tier that does not reflect their actual level of achievement in that subject. On the other hand we have found little support for the ambition to do away with tiering in all EBC subjects. There are strong grounds for supporting the view that differentiation by outcome is suitable only in some curriculum areas. It is very hard to see how appropriate single common paper approaches can be successful in subjects such as mathematics. As stressed by many assessment experts and respondents to the DfE's consultation, pulling together examination questions of varying difficulties within a single three hour examination in order to try to meet the needs of candidates of widely differing levels of achievement will greatly increase the risk of mis-grading. It will cause such examinations to have dangerously low levels of validity and reliability.

8.6 Perhaps more important than these grading issues are the broader questions involved in setting examinations that are fair, valid and equitable. Most important of all is the encouragement of young people to achieve the highest levels of learning of which they are capable. As we have noted, there is a concern that tiered examination papers may cap the aspirations of some learners. The evidence from this review is far less clear on that point, and specifically highlights the very real emotional, educational and social consequences of asking whole age cohorts to attempt new EBC examinations, which may in the event be quite unsuited to their current levels of achievement in a wide range of educational areas.

8.7 It is the case, as we have noted, that some research studies have revealed problems with the way that students are allocated to different tiers both in GCSE examinations here and in similar situations in other countries. Responding to such evidence by arguing that tiering itself is the problem is, we have suggested, misplaced. Those responsible for developing the models for EBC examinations should consider both the strengths and weaknesses of all approaches to differentiated

examinations at the end of Key Stage 4, and consider how teachers can best be assisted in adopting the very best classroom practices to support both teaching and assessment in the implementation phase of the new EBC qualifications. As ever, what is seen by some as a simple assessment problem is in fact a complex pedagogical challenge for both teachers, examiners and ministers, as together they seek to inspire learning and achievement among a diverse group of learners.

Appendix: Selected responses to the DfE Consultation on Reforming Key stage 4 Qualifications (2012)

Reforming Key Stage 4 Qualifications: Consultation Response Form (Department for Education) **5 Do you agree that it will be possible to end tiering for the full range of subjects that we will be creating new qualifications for?** **6 Are there particular approaches to examinations which might be needed to make this possible for some subjects?**	
Respondent	**Response**
Wellcome Trust	We believe that it will be extremely difficult to assess the necessary range of student achievement without the use of tiered examinations. It is inefficient and ineffective to try to discriminate the full range of abilities of students with a single examination. Tiered examinations would enable EBCs to be adaptively delivered to a greater range of students. We are highly concerned by the lack of explanation of how students with practical and creative rather than academic strengths will have those recognised in a way that is valued by them, their parents and their future employers. Failure to tackle this deficit will create a legacy of demotivation that is all too familiar from the days of O level.
Science Community Representing Education (SCORE)	We are also unconvinced that a single tiered assessment system can accurately and fairly measure achievement across all abilities whilst raising aspirations and increasing the motivation of all. The early years of GCSEs existed in a three tiered system (foundation, intermediate and higher) which allowed for a greater degree of stretch and challenge in the assessment items, while also allowing for lower-achieving students to receive a grade demonstrating their level of achievement. Moving from a three-tiered to a two-tiered model has made it harder to differentiate between the most able students because there are not enough questions at the A/A* level. A single tiered system would exacerbate this problem, particularly if one of the purposes of the assessment was to provide certification: • It would fail to enable students at either end of the achievement scale to demonstrate their abilities effectively, resulting in question papers that are off-putting for those that find it too difficult and trivial for those that find it too easy. It would thus be unlikely to meet one of the consultation's main aims, which is to provide differentiation for strong performance and recognition of lower levels of performance. • Assessments would also have to be very long in order to allow for sufficient items to assess the full range of ability. There are other models of assessment that could be investigated to address this issue, for example a paper to demonstrate competency in a subject taken by all, plus an additional paper taken by high achievers to provide differentiation. SCORE recommends that the Department for Education consults with assessment experts to explore these possibilities.
Chartered Institute of Educational Assessors (CIEA)	In order that valid papers can be set for almost the full ability range, it is important that the concept of tiering is retained. In subjects which rely heavily on interpretation and professional judgement, the removal of tiering is likely to make it more difficult to ensure precision and thus public confidence.

Association of Directors of Children's Services	Tiering should be considered on a subject basis, by curriculum experts making judgements about the utility of lower tiered papers in providing a coherent curriculum for those who might not obtain the highest grades. In some subjects offering firm foundations of basic concepts will be more productive than trying to cover the full curriculum more thinly. However we do agree that limiting the lower tier papers to a grade C, rather than the overlap to a grade "B" as in maths, does limit ambition. The lowest tier papers should allow for a "B" grade.
Mathematics in Education and Industry (MEI)	Ending tiering may be appropriate for some subjects, but it would not work for mathematics. There is considerable evidence that the range of achievement in mathematics at age 16 is very wide. Furthermore, mathematics assessment tends to differentiate by question, rather than by outcome; questions that mathematically able students find easy can seem impossible to average students, whereas the essay-style questions commonly used in other subjects can stimulate a full range of responses from students of different abilities. We need a mathematics qualification that can challenge the mathematically-able, whilst also providing assurance of numeracy for all. Without tiered assessment, it is not possible to meet the needs of all students in mathematics.
Royal Historical Society	We are clear in our opposition to tiering in our own subject of History, in which tiering does not currently operate. We would not be pleased to see it introduced in History in order to secure uniformity of practice across disciplines, as we agree with the assertion in the consultation document that it would unnecessarily 'cap' student performance in a qualification in which achieving the highest possible grade will be of considerable importance to a candidate's future prospects, especially in the absence of the opportunity to resit papers. While we are not convinced it is essential in any other discipline, we would be content to see other subject communities who have different views and can make a specific case for its continuation to be allowed tiering in any new qualification. We would, however, still have concerns that the practice of tiering may act against the interests of some students whose developmental patterns would be ill served by a rigid division of students into different pathways at an early stage. It is important to have 'open' questioning which can allow all candidates to answer a common question at a variety of different levels, demonstrating the level of their ability in the quality of the answer, rather than in a decision on whether to attempt a question or not.
Geographical Association	It is possible to have common papers in geography but it is necessary to recognise that this will result in significant technical assessment challenges for examiners and schools. In the past some Geography GCSE papers were common to the whole cohort and while this removed barriers to achievement in the B/C/D grade area it had the potential to cause problems for lower achieving candidates who found themselves unable to answer many questions, whilst the highest-attaining candidates were neither challenged nor identified by some sections of the same papers. Differentiated papers in geography are advantageous because the same content can be assessed in different ways and questions can be structured to enable all candidates to demonstrate what they know understand and can do in appropriate ways. Spending time in an examination where all the questions are too difficult can lead to frustration and demotivate students. To overcome problems like this it might be necessary to employ new assessment approaches such as differential time allowances, or provide an element of additional challenge to the very highest attainers. The GA also regards this review as an opportunity to think afresh about more creative assessment methods, which allow variety whilst ensuring comparability... Geography examination questions have usually been structured with an incline of difficulty where earlier questions are more guided and later questions are more open ended. In general this structure has proved successful because it enables differentiation in

	a number of ways and provides opportunity to include a variety of question styles within the same examination. Knowledge, understanding and skills can be identified and assessed. More open ended questions allow differentiation by outcome because candidates can answer in a variety of different ways. There are however also known disadvantages. Guided questions can disadvantage some high performing candidates who either spend too long on less valuable questions or attempt to make their answers too complex and sophisticated. The specificity of guided questions can lead to some lower achievers underperforming because they have insufficient breadth of knowledge, understanding and skills. Open ended questions can lead to simple regurgitation of 'all the candidate knows about the topic', rather than the anticipated response. In a comparison of geography and history GCSE questions conducted by QCA in 2008 it was found that while, superficially, geography questions appeared 'easier' because they were more structured, in reality student performance in the two subjects was largely comparable. It was also found that the content of geography examinations were less predictable so that students were required to know a wider range of content from the syllabus.
University Council of Modern Languages	Grade capping through tiering is not a useful tool and candidates should be given chances to outperform expectations which are currently capped by decisions on the level at which they are entered. Languages: examinations should be structured to differentiate performance by output. Currently, candidates entered at foundation level only may not have the opportunity to demonstrate some aspects of language learning only tested in higher level exams. We would like to see all pupils encouraged to cover the whole curriculum and to be given the chance to demonstrate the skills they have achieved, through inclusion of more open-ended question sets.
Advisory Committee on Mathematics Education (ACME)	ACME is very concerned about the Government's proposals to eliminate tiering in mathematics. Whilst we understand that the current tiering framework has on occasion been misused to limit aspiration, we have grave reservations about the impact of having a single paper for all grades in mathematics. Mathematics assessment is different to that in other subjects as it differentiates by task and not by outcome. In other subjects, the questions can be the same for all students; the mark scheme indicates how different responses should be evaluated. In mathematics examinations, on the other hand, questions at quite different standards are set for candidates of different attainments. In mathematics especially, there is a very broad range of attainment at Key Stage 4. For these reasons, a single paper for all grades is impracticable, as weaker candidates would be faced with challenging questions they could not begin, and strong candidates would waste time working through questions that present them with no challenge whatsoever; this would also make for an exceptionally long paper. It would therefore be very difficult to design a valid examination which would reliably assess the learning of all, or even 90%, of students at age 16. Finally, there is currently no expertise in Awarding Organisations' mathematics teams in how to set and mark such a paper so that it delivers reliable outcomes across the grade range. Consequently, there is a broad consensus within the mathematics community that having a single paper for all grade ranges is neither feasible nor desirable in the immediate future. An alternative proposal that could address the issues outlined in the consultation would be that of a three-level system, where students are entered for two adjacent sets of papers. The Key Stage 4 qualification might be assessed at say Foundation level, Standard level and Higher level. All students would sit the Standard papers and either the Foundation or

	Higher paper. Thus all students would have examinations where they could demonstrate positive achievement, as well as being stretched. This system would also allow students to sit common Standard level papers which would permit comparison to be made between all students. Each level would need to test both methods and application aspects of the subject. ACME believes that in order to maximise each student's potential to achieve at the highest level, it would be crucial that the final decision about which level would be sat by each individual be left until as late as possible in KS4. Students develop at different rates and, in the short-term, many KS3 students will continue to be taught by non-specialists; for some young people, KS4 will be the first opportunity to experience specialist teaching. In addition, in order to ensure that such a model does not put a cap on expectations, it should ensure that a student who did very well in Foundation and Standard papers should be able to bridge to A level mathematics.