Moving to two-tier GCSE mathematics examinations

An independent evaluation of the 2005 GCSE Pilot and Trial

Gordon Stobart, Tamara Bibby and Harvey Goldstein

University of London Institute of Education

with the assistance of
Ian Schagen (NfER), and
Mike Treadaway (Fischer Family Trust)

September 2005

Acknowledgements

This evaluation was conducted over a very short time period and would not have been possible without the full cooperation of the awarding body researchers and mathematics subject officers. We thank them for providing data at a very hectic time in the examination cycle. Particular thanks go to Eddie Wilde and Mike Forster at OCR who had to provide multiple data files.

We are particularly grateful to Ian Schagen (NfER) who conducted the statistical analyses of the examination data to very tight deadlines. Thanks also to Mike Treadaway (Fischer Family Trust) for his matching of key stage 3 data to GCSE candidates.

The help we have received from QCA colleagues, particularly Tina Isaacs and Colin Robinson, has also facilitated this evaluation

Executive Summary

This independent evaluation was commissioned as part of QCA's work to assess the relative merits of the two schemes of 'two-tier' GCSE mathematics which were examined, alongside the current three-tier model, in the summer 2005 examinations.

These two-tier schemes, the OCR Pilot and the AQA/Edexcel/OCR/WJEC Trial, are similar in key respects:

- they involve the same curriculum;
- both Foundation and Higher tier candidates can achieve grades C and D;
- the coursework contribution is the same.

The key differences are:

- the degree of overlap between the papers;
- the grading process.

It the impact of these differences that is the focus of this evaluation.

1.1 The Pilot model

The Pilot model is in the third year of awarding grades. It is, for GCSE, a unique 'two tier' model in which candidates take two adjacent papers, each covering a limited range of grades (A*-B; C-D; E-G) from which the best grade is selected. The marks from the two papers are not combined.

1.2 The Trial model

This is the 'standard' two tier model which operates in all other 'tiered' GCSE subjects. The tiered papers cover overlapping grades (Higher A*- D; Foundation C-G) and the data from both papers are combined to arrive at the final grade.

1.3 Key Issues

These differences raise some key questions which are central to this evaluation:

i. What are the main purposes that the GCSE mathematics examination serves?

The central tension here is between general certification of over 90% of the 16 year old cohort at the end of compulsory schooling and the identification of a minority of students who may progress to further study of mathematics. The Tomlinson Report (2004) discussed such tensions in terms of *inclusion* and *stretch.*

Our evaluation considers the relative accessibility of the two models, particularly student and teacher perceptions of the examinations.

ii. What does a GCSE grade mean and how is it best determined?

This question looks at the information a grade carries about a student's competence in mathematics and the reliability of this information.

Both the Pilot and Trial models are 'compensation' models in that a grade is determined by the aggregated marks rather than by meeting specific performance criteria (Pilot: aggregated within papers, Trial: aggregated across papers). The Pilot has sought to target more narrowly the content and skills which are tested on each paper. We examine how effective this targeting was by looking at the level of demand of the papers set.

iii. How reliable are the grades awarded?

We looked at the comparability of the Trial grades with the candidates' grades on the current three tier examination or the Pilot. We also analysed whether papers and questions behaved as intended (eg were they at the intended level of difficulty?).

2.1 The QCA brief

The brief provided a series of issues and the expectation that analyses of the examinations would be complemented by questionnaires, in-depth interviews, focus groups of students and teachers, and centre visits.

2.2 The 'performance' data

We collected data, with the help of the awarding bodies, for the Trial and three-tier examinations. OCR also provided data for the eight schools in which students took both the Pilot and Trial examinations. Each awarding body also provided a sample of question level marks. We also had access to the key stage 3 data of the candidates.

The OCR Pilot and OCR Trial awarding meetings were attended.

2.3 The 'perceptions' data

We collected survey data from 76 schools. 93 teachers and 1575 students returned questionnaires on the Trial. Ten teachers from five schools of the schools taking both the Trial and Pilot responded to a separate survey, as did 430 students. We also visited 8 centres and interviewed 31 teachers and 60 students.

Main findings

Accessibility and inclusion vs. demand.

The impact of the different paper structures on the students taking them is important as it may affect their exam performance and colour future attitudes to mathematics. Our survey findings showed that **overall 60% of Trial students preferred it to the three-tier examination** – even though they had had little or no time to prepare for it.

We had survey returns from 430 students (in five schools) who sat both the Pilot and Trial examinations. **Two thirds of these students preferred the Trial to the Pilot model.** The reasons they gave for this were generally related to the range and accessibility of the questions and the confidence-building effects of easier lead-in questions (Table 6.2).

The mathematics teachers in the schools that took both the Trial and Pilot examinations were divided in their preferences. The split was between more challenge for higher achieving students plus more targetted teaching of specific topics (Pilot preferred) and those who saw the Trial as a more positive experience for the students with the promise of better teaching.

The most consistent negative comment about the Trial papers was the time allowed (2 hours per paper), which both teachers and students thought was too long, as most candidates finished well inside this time.

The Trial grade distributions were generally slightly lower across each of the awarding bodies and the OCR Pilot than the students' three-tier results across all the awarding bodies. To compensate for this our analyses brought the overall Trial grade distribution into line with both Pilot and 3-tier examinations.

The comparison of the adjusted **Trial and three-tier grade distributions showed that only around 70% of candidates received the same grade on both examinations**

When the adjusted Trial and Pilot results were compared to the 'fine grade' KS3 results, we found that **for those candidates who had achieved levels 2-6 at key stage 3, the Trial offered a better chance of a higher grade. It was only for students gaining level 7 or above that the Pilot was likely to confer some advantage although most of these candidates did equally well on both models** (Figure 3e).

While we need to be cautious, **there is evidence that a move to a 2-tier examination structure will encourage girls to continue studying mathematics at Alevel** (Table 7.5).

What does a GCSE mathematics grade mean?

We found on the Pilot that our question level data challenged the assumption that particular questions could be precisely targetted at particular grades (Sections 3 & 4). The question level data from the AQA Trial showed similar variability, though is less dependent on the precise targeting of questions.

What information does a grade carry?

The grade B on the Trial allows very few inferences about a candidate's particular mathematical knowledge or skills, since the overall marks can be achieved by multiple routes. The Pilot claims to carry more grade-related information about the skills and knowledge**. Our analyses challenge the assumption that questions can be precisely targetted at particular grades.**

For the Trial model there remain two routes to grade C, which raises well rehearsed comparability issues.

Implementation issues

Teachers did see not moving to a two tier model as particularly problematic in terms of classroom organization, setting or resources.

Teachers were most negative about GCSE coursework which they believed to be time-consuming and increasingly unreliable because of the ease of accessing model answers.

Conclusion

The move to a two-tier GCSE mathematics examination has been widely welcomed by teachers and students. The two models we have evaluated have much in common in terms of curriculum and the accessibility of the iconic C grade for all candidates. If one model is to be chosen, we believe the key questions are about the main purposes that the GCSE mathematics serves and what can be inferred from a grade. We see a different emphasis in relation to inclusion and stretch between the Trial and the Pilot. We have questioned, for each model, what information a grade carries. Our evidence suggests that grade based inferences about mathematical competencies have to be extremely cautious, particularly as targeting the difficulty level of questions and papers proved difficult for both models.

The strength of the Trial is its accessibility, that of the Pilot is the apparent stretch provided, particularly for those who gained level 7 in their key stage 3 tests. The obverse of this provides the risks: the Trial provides lower demand for the highest achievers, and the structure of the Pilot means that for many students one paper is inaccessible.

Contents

List of tables and figures

1. Overview

This independent evaluation was commissioned as part of QCA's work to assess the relative merits of the two schemes of 'two-tier' GCSE mathematics which were examined, alongside the current three-tier model, in the summer 2005 examinations.

These two-tier schemes, the OCR Pilot and the AQA/Edexcel/OCR/WJEC Trial, are similar in key respects:

- they involve the same curriculum;
- both Foundation and Higher candidates can achieve grades C and D;
- the coursework contribution is the same.

The key differences are:

- the degree of overlap between the papers;
- the grading process.

It the impact of these differences that is the focus of this evaluation.

1.1 The Pilot model

The Pilot model is in the third year of awarding grades. It is run by the OCR awarding body on behalf of QCA, which was directly involved in its design. It is, for GCSE, a unique 'two-tier' model in which candidates take two adjacent papers, each covering a limited range of grades (A*-B; C-D; E-G, see Fig. 1) from which the best grade is selected. *The marks from the two papers are not combined,* so the information from one paper is ignored. There is no overlap between papers and there is only one route to achieving each grade, for example a grade C can only be attained on the middle paper - which all candidates take. The assumption is that questions can be accurately targeted at specific grades so that, for example, all the questions on the middle paper will be at grade D or C level of demand in terms of content and difficulty.

Each paper is two hours long and is evenly split between calculator and non-calculator sections. The weighted mark from the coursework component is incorporated to generate the final grade. Pilot students did not take the three-tier examinations, though in eight schools they also took the Trial papers in 2005, with the best result being certificated.

1.2 The Trial model

Concerns at some of the data which emerged from the 2003 and 2004 Pilot evaluations and experience with the 'standard' two-tier model which operates in all other 'tiered' GCSE subjects led to pressures, primarily from the awarding bodies, to trial an alternative scheme. QCA approved this for the summer 2005 examinations.

This timescale has meant that candidates had been prepared in Years 10 &11 for the current three-tier examination, with selection for the Trial only being decided in March/April 2005. This resulted in teachers and students having limited opportunities to prepare for the wider range of questions that would be found on the two-tier papers. This lack of preparation confounds any interpretation of results since it may not be easily possible to disentangle 'start-up' factors (e.g. topics not covered in class) from possible 'structural' problems (e.g. differentiating across five grades). Another factor may be that three awarding bodies'

students sat the Trial as their very last GCSE examination, sometimes a week after any other GCSEs. Only AQA students sat the Trial before the three-tier examination. The candidates sat both the three-tier and the Trial examinations with the best final grade being certificated.

The salient features of the 'standard two-tier' Trial model are that papers cover overlapping grades (A*- D and C-G, see Fig. 1) and the data from both papers are combined to arrive at the final grade. Candidates were entered for either the Foundation or Higher tier and sat two two-hour papers, one with and one without a calculator. Raw marks from each were converted onto a uniform mark scale (UMS), as were the coursework marks, and combined to generate the final grade. There were two routes to grades C and D and there were some common questions at these grade levels across the two tiers to help with grading comparability.

Figure 1: Models of tiered GCSE mathematics examinations

		A*	A	B	C	D	E	F	G	U
'Traditional' 3-tier Students entered for appropriate tier and sit two papers at that level. Scores are combined and averaged to give grade.	Higher	■	■	■	■					
	Intermediate			■	■	■	■			
	Foundation					■	■	■	■	
'Stepped' 2-tier (QCA/OCR Pilot) All students sit 'core' papers and *either* the Higher *or* Foundation papers. Students gain the grade from the highest paper. Scores from different levels of paper are not combined.	Higher	■	■	■						
	Core				■	■				
	Foundation						■	■	■	
'Overlapping' 2-tier (2005 Trial) Students entered for appropriate tier. Scores on papers are 'averaged' to give overall grade.	Higher	■	■	■	■	■				
	Foundation				■	■	■	■	■	

The 'traditional' 3-tier model is included for comparison ■ *Levels normally available.*

1.3 Key Issues

These differences raise three key questions which are central to this evaluation:

i. What are the main purposes that the GCSE mathematics examination serves?
The central tension here is between general certification of over 90% of the 16 year old cohort at the end of compulsory schooling and the identification of a minority of students who may progress to further study of mathematics. The Tomlinson Report (2004) discussed such tensions in terms of *inclusion* and *stretch.*

In GCSE mathematics which, more than any other GCSE subject, links specific content to particular grade levels, this tension is particularly acute. Inclusion is limited if many students are not taught the content/skills needed for a higher grade. Stretch is reduced if the examinations emphasise the content and skills which the majority have covered. In terms of this tension we see the Trial structure offering a more inclusive approach while the Pilot seeks more stretch. We do not think the current structure of GCSE mathematics allows both purposes to be fully met.

Our evaluation considers the relative accessibility of the two models, particularly student and teacher perceptions of the examinations, by asking:

- How positive was the experience of the different mathematics examinations for the candidates?

- Given the influence of examinations on student attitudes and motivation, what was the impact of each model?

- What were the perceptions of the two examinations of those students who may consider further study of mathematics?

ii What does a GCSE grade mean and how is it best determined?

This question involves the information a grade carries about a student's competence in mathematics and the reliability of this information.

What can be inferred from a GCSE mathematics grade?

If students achieve a particular grade, what does this tell us about their mathematical knowledge and skills? This is a particularly sensitive question at the higher grades since there is an assumption that these students – who may move on to GCE A level mathematics courses – will have developed skills in, for example, algebraic manipulation.

Both the Pilot and Trial models are 'compensation' models in that a grade is determined by the aggregated marks rather than by meeting specific performance criteria. In either model, if a grade boundary is 40 marks it does not matter how the student gets these marks. The concern is that students may get a grade A or B with little or no success on questions targeted at this level since they collected their marks by scoring highly on less demanding questions. For example, in the context of the current three-tier GCSE, selectors will often require a grade B from the Higher tier, rather than from the Intermediate tier, for admission on to AS mathematics courses. The assumption is that Higher tier students will have been exposed to higher level content than Intermediate grade B students.

The Pilot examination has sought to lessen the risk associated with multiple ways of achieving a grade by restricting the range of demand on each paper. To get a grade B a candidate must take the A*-B paper, on which the questions reflect higher level content and skills. The assumption here is that questions can be targeted accurately at a particular grade in terms of their accessibility and demand. In relation to content, this confidence is based on topics being grade related, for example only those candidates predicted grade A*- B will be taught more complex trigonometry, vectors and probability. There may be a self-fulfilling element in this – a candidate who, based on KS3 results, is predicted a grade C is not expected to master this content so is not taught it – which leads to the expected grade C.

Our evaluation focuses more on the *level of demand of the questions set* – did they behave as intended? For example, did the questions on the Pilot middle paper (grades C & D) all prove easier than those on the higher paper (grades A & B)? Were grade boundaries set around the target mark (for example grade A at 65-75% of the raw marks)?

The Trial questions do not assume quite this level of precision, with examiners working with three levels of demand (low, medium and high), though there were still grade related weightings of questions (for example 20% of the marks on the higher tier were targeted at grade A*/A/higher levels of demand). The assumption is that the grade represents the overall attainment across the five grade levels, and broader content base, of each tier. A concern is that little in terms of specific content and skills can be inferred from a particular grade, with grade B again problematic for progression to AS work.

iii How reliable are the grades awarded?

Our evaluation looks at several reliability concerns. One is whether the Trial produced similar distributions of grades to the thee-tier examination and the Pilot for students who took both. This is essentially about comparability should one system replace the other. Even if the overall pattern is the same will some groups benefit and others do worse as a result of the change?

A second concern is whether the examination papers behaved as intended. Did the grade boundaries fall in the intended mark zone, for example, a grade A in the 65-75 per cent of raw marks range? If the distribution of raw marks is very different to the intended one, for example selecting a grade boundary on a very low mark, what are the implications for grade reliability?

A third, and key, concern is whether the questions on the papers were of the level of demand intended. This is critical for the Pilot as the intention is to narrowly target the level of demand of the questions. If the questions did not function as intended this impacts on reliability as a grade may be determined by relatively few questions (since the information from the other paper is not used).

We address these concerns through our analysis of the examination data collected for the evaluation.

1.4 The structure of this report

In responding directly to the QCA Brief, we have provided little of the more general background to these developments (e.g. Smith, 2004). We focus on the two main strands of work that were undertaken.

Performance. This involves the collection and analysis of Trial, Pilot and three-tier examination data which was in turn matched to candidates' key stage 3 scores. This allowed further modeling of relative progress and school effects. The awarding bodies (AQA; Edexcel; OCR; WJEC) also provided detailed question level data on a sample of students which allowed analysis of how they responded to particular questions. Coupled with this is an account of the OCR award meetings for both the Pilot and Trial examinations which summarises the awarding procedures and the awarding issues each model raised.

Perceptions. This strand focuses on teacher and student perceptions of the Trial and Pilot examinations. These provided insights into the preparation for, and response to, the two-tier examination. While many Trial students and teachers were responding in relation to the three-tier examination for which they had been prepared, we also discuss the responses of students who took both the Pilot and Trial examinations.

We present these as separate sections before drawing together the main findings. We have sought to keep the main report accessible by presenting much of the detailed analysis in the Appendices.

2. Methods

2.1 The QCA brief

The brief provided a series of issues to be discussed (see below) and the expectation that analyses of the examinations would be complemented by questionnaires, in-depth interviews, focus groups of students and teachers, and centre visits.

Relating to performance of the examinations

- Reliability, validity and manageability of the assessment regimes investigated (including any teacher assessed elements) and of the grading and awarding system;
- Comparability of grade standards within and between awarding bodies;
- Comparability of grade standards between a new assessment model and the current three-tier assessment model;
- Positioning of grade boundaries and the public credibility

Relating to teacher perceptions

- Teachers' attitudes to, and confidence in, teaching the content and approaches to mathematics in the three-tier, Pilot model and Trial model;
- Teachers' attitudes to, and confidence in, undertaking different approaches to assessment included in the pilot, and anticipated continuing development demands of this;
- Issues for schools and teachers, including professional development needs, staffing, timetabling, manageability, resourcing and links with the local community;
- Issues for awarding bodies in terms of, for example, examination timetabling, examiner supply.
- The effects of the proposed examination structure on teaching and learning programmes;
- The effects of the proposed examination structure on progression from GCSE to advanced level study in mathematics;

Relating to student perceptions

- Students' attitudes to, and experience of, the content, questions and approaches of the examination;
- Students' attitudes towards their own progression in mathematics;

In addition to the data we collected for this evaluation, various documents were made available to us by QCA, OCR and AQA. These included information about the development of both models, evaluations of the 2003 and 2004 Pilot examinations, and the response to these from AQA. We have also talked to awarding body personnel and QCA staff, including the mathematics team. We also got feedback on our approach from a meeting with the Advisory Committee on Mathematics Education.

2.2 The 'performance' strand

We collected data, with the help of the awarding bodies, on entries for the Trial and three-tier examinations. OCR also provided data for the eight schools in which students took both the Pilot and Trial examinations. This was then matched, after the awarding meetings in July, to

performance data (paper marks, overall marks and grades etc.) on both examinations. We also asked each awarding body to provide a sample of question level marks for around 400 students drawn from around 10 schools. This would allow detailed analysis of performance on specific questions.

We also had access to the key stage 3 data of the candidates and this was matched at with individual candidates to allow multilevel modeling of progress to investigate how key stage 3 scores were related to the different examinations.

The Award Meetings

Both the OCR Pilot and OCR Trial awarding meetings were attended. These were each two-day meetings in Cambridge in July. The intention was to assess similarities and differences in the procedures and to discuss with examiners what issues the different models raised.

2.3 The 'perceptions' strand

We sought the responses of students, teachers and examinations officers, from a variety of schools and across each of the exam boards.

The Sample (see Table 2.1 below)

Different questionnaires (teacher and student) were sent to those taking the 3-tier and Trial papers, and those OCR schools taking the two 2-tier Pilot and Trial papers. Copies of all four questionnaires are included in Appendix 1)

Table 2.1 Questionnaires sent out to, and received from, schools

Exam board	**3-tier & Trial**					**Pilot & Trial**				
	Schools		Teacher	Student		Schools		Teacher	Student	
	sample	returned	returned	out	*returned*	sample	returned	returned	out	*returned*
AQA	14	10	28	934		-	-	-	-	-
Edexcel	29	15	35	490		-	-	-	-	-
OCR	13	5	12	221		6	5	10	623	430
WJEC *	15	6	18	186		-	-	-	-	-
Totals	**70**	**36** (51%)	**93**	**1831**	**1575**	**6**	**5** (83%)	**10**	**623**	**430**

- In all 76 schools took part in the evaluation.
- Every school was sent 6 teacher questionnaires.
- All students in each of the six schools where the students took both the Pilot and Trial papers were sent student questionnaires.
- A sample of four schools per exam board in the 3-tier and Trial groups were sent student questionnaires – these were chosen to reflect a variety of sizes and types of schools in different geographic locations. Where possible we selected schools with a balance of entries to 'Foundation' and 'Higher' tiers.
- A sub-set of two schools per exam board (apart from WJEC where only one school was visited due to logistical difficulties) were then selected and visited on the day of the final

mathematics examination – two small groups of students were interviewed as were staff who had taught groups involved in the Trial/ Pilots.

- A small sample of schools were contacted by telephone after the exam results had been given out to gauge responses to the results.

Table 2.2 Interviews conducted

Exam board	Center visits	Student interviews	Teacher interviews	Post exam result telephone interviews
AQA	3 **	22	9	1
Edexcel	2	15	7	1
OCR	2	19	9	2
WJEC	1	6	6	
Totals	**8**	**60**	**31**	**4**

** includes visits to trial interviews and questionnaires

3. Analysis of examination data

This section summarises analyses of the data provided by the awarding bodies and key stage 3 data for these GCSE candidates. We address three questions:

1. How did the results of the Trial match the candidates' three tier results?
2. What was the match for those candidates who took both the OCR Trial and OCR Pilot and how do their results relate to their key stage 3 scores?
3. What does the question level analysis reveal about the actual difficulty of questions in relation to the intended difficulty?

3.1 How did the results of the Trial match the candidates' three tier results?

This analysis compared paper and UMS scores and final grades awarded. Data were available for Trial candidates from each awarding body. Initially we focus on the comparisons between the grades awarded. Table 3.1 summarises the overall levels of agreement of candidates' Trial and three tier examination grades for each awarding body. We report both the overall measure of agreement (Kappa) lying between 0 and 1 and the average grade difference.

Table 3.1 Trial entries

Awarding Body	Number of candidates	% of students getting same grade from both exams	Kappa (adjusted)	Grade difference (3-tier – Trial)
AQA	3437	62.3	0.55 (0.60)	0.27
EDEXCEL	2773	72.7	0.66 (0.71)	0.15
OCR	1203	57.1	0.49 (0.61)	0.34
WJEC	2184	68.6	0.63 (0.65)	0.03

Further details are given in Appendix 2.

Given that the overall grading on the Trial was slightly lower that on the 3-tier, it is not surprising that the level of exact matches was only modest. In order to take account of this we adjusted the 3-tier and Trial grades to have the same (marginal) distribution so that overall differences are eliminated. This was carried out by using the total UMS score for the Trial papers, developing 'cut-scores' (maximum UMS scores to be awarded each grade) in order to match as closely as possible to the distribution of grades achieved on the 3-tier examination. This increases the measures of agreement slightly to the ones given in brackets. Part of the lack of complete agreement is due to the inherent unreliability of grading and we have no independent estimate of this. Nevertheless, it does seem that moving from a 3-tier

system to a Trial system could, even if the overall grade distribution is similar, produce different individual results for a number of candidates.

These levels of agreement are modest and Figure 3a summarises the grade differences over all awarding bodies for each grade on the 3-tier examination. Thus, for example, for the 3-tier unclassified (U) candidates approximately 25% are also unclassified on the Trial, about 10% obtain 1 grade higher and about 65% obtain grades two or more higher. Care needs to be taken interpreting results relating to candidates obtaining a U grade on the Trial, however, since this was a voluntary examination taken at the end of the examination period so, for example, students may not have turned up for both papers.

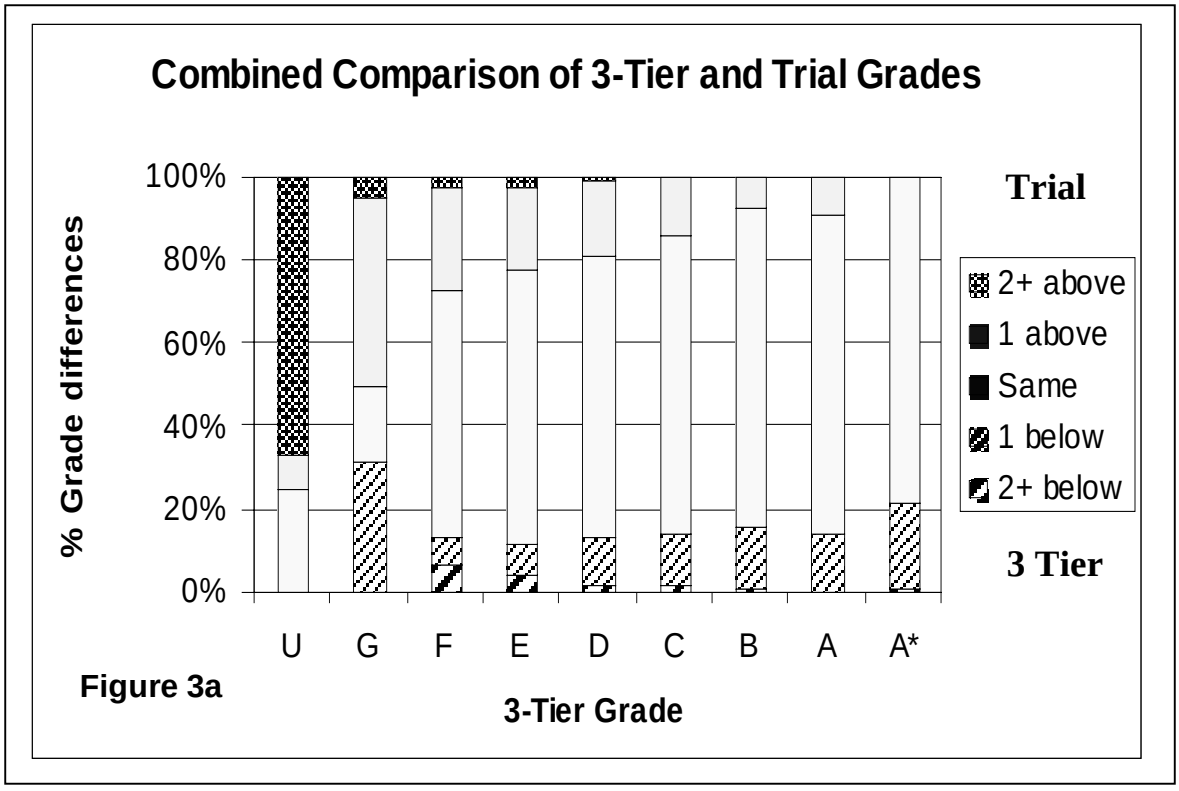


Figure 3a

3.2 What was the match for those candidates who took both the OCR Trial and OCR Pilot and how do their results relate to their key stage 3 scores?

A total of 740 candidates were entered for both the Pilot and Trial papers in OCR. On average the Pilot grades were 0.82 grades higher than the Trial ones. When U grades were omitted this became 0.34 grades. The distribution of grades is shown in Table 3.2

Table 3.2 Distribution of candidates' Trial and Pilot grades

	Pilot Grade									
Trial grade	U	G	F	E	D	C	B	A	A*	Row total
U	10	15	19	7	25	28	12	1		117 15.8
G	1	3	23	4	1					32 4.2
F			13	20	9	2				44 5.9
E	1	1		17	63	3				85 11.5
D				2	49	50				101 13.6
C					5	110	40			115 20.9
B						9	95	8		112 15.1
A							7	61	2	70 9.5
A*								8	16	24 3.2
column total	12 1.6	19 2.6	55 7.4	50 6.8	152 20.5	202 27.3	154 20.8	78 10.5	18 2.4	740 100

What Table 3.2 demonstrates is that only half the candidates got the same grade on both examinations. The lower grading of the Trial meant that only 34 candidates improved their grades on the Trial, while 332 got a higher grade on the Pilot. This includes a 107 candidates who were Unclassified on the Trial but gained better grades (31 grade C and above) on the Pilot. This may have explanations we could not infer from the data.

Using combined paper scores in the Pilot

A concern about the Pilot model is whether discarding the marks of one paper may reduce the reliability of the award since it will be based on less information. The alternative approach, adopted in the Trial, is to use the total UMS score on both elements to define the final grade. We looked at what difference it would make if the Pilot used the marks from both papers. From the Pilot data, the total UMS 'cut-scores' to define grade boundaries were set in such a way as to give the same distribution of grades as were actually awarded in the Pilot. These 'pseudo-Trial' grades were then compared with the actual Pilot grades (for all candidates taking the Pilot). What this showed was relatively strong overall agreement, with a Kappa value of 0.93, and with 95 per cent of candidates getting the same grade (see Appendix 3). This still means that one in twenty candidates would have got a different grade by this method (if extrapolated to the GCSE cohort some 35,000 students). Of the 368 such Pilot candidates, (out of 7420), 80 were grade C candidates who moved up to grade B, while a further 97 moved from a B to a C. Given the low grade boundary for grade B (24%), which can be seen as a threat to reliability, this combination could be considered as a means of

improving reliability. No grade A or A* candidates changed grades, while 70 F and G candidates would have gone down a grade.

How do candidates Pilot and Trial results relate to their key stage 3 scores?

A total of 730 candidates had results for both the OCR Pilot and Trial examinations, as well as KS3 'fine grade' results for core subjects. These candidates were used as the basis for a detailed investigation of the relative probabilities of achieving different grades, conditional on KS3 mathematics performance. The 'fine grade' score uses decimals to split each level into 10 sub-levels covering all the levels so, for example, 'level 5.3' can exist in this format.

The first step was to ensure that the distributions of grades on the two exams were the same, so that we could eliminate effects due to any systematic differences in grading and focus on differences due to the structures of the two tests. This was done by examining the cumulative distributions of the Pilot grades and the Trial UMS marks, and finding values of the latter which corresponded as closely as possible to the former. These points were then used to define 'rescaled' Trial grades with essentially the same distributions as the Pilot grades.

For each examination, for each grade, the probability of achieving that grade or higher was related to the KS3 mathematics fine grade score (using a multilevel logistic model) The results are summarized in Appendix 4. For ease of presentation we look only at grades F, C and A (Figures 3b-d).

For the probability of obtaining a grade F or higher we see that the Trial candidates with low KS3 scores do much better (Figure 3b).

For the probability of obtaining a C or higher the same is true but only up to about the median KS3 score, after which there is an advantage for the Pilot (Figure 3c).

For the probability of obtaining an A or A* there is little difference between the Pilot and Trial (Figure 3d).

Figure 3b

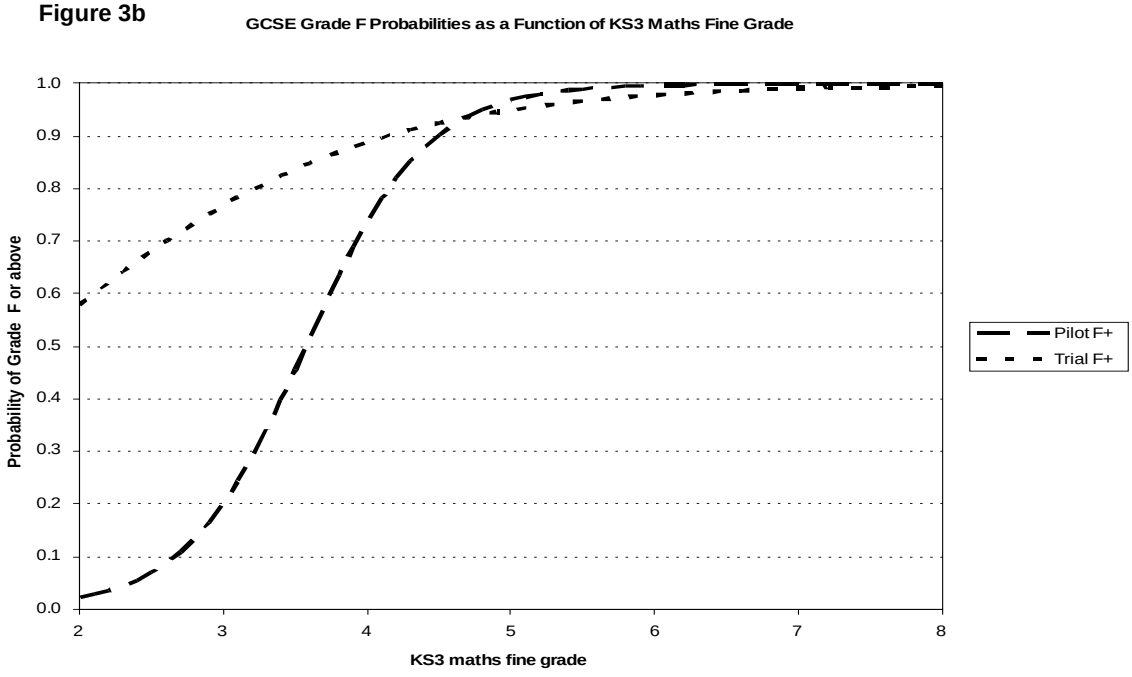


Figure 3c GCSE Grade C Probabilities as a Function of KS3 Maths Fine Grade

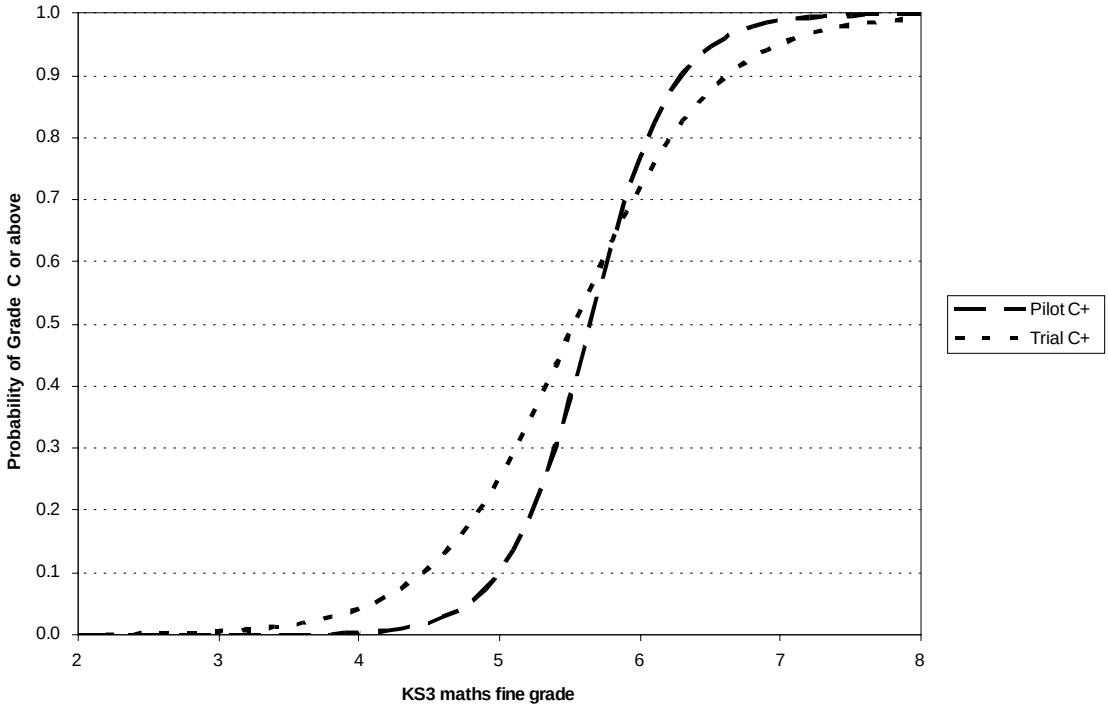
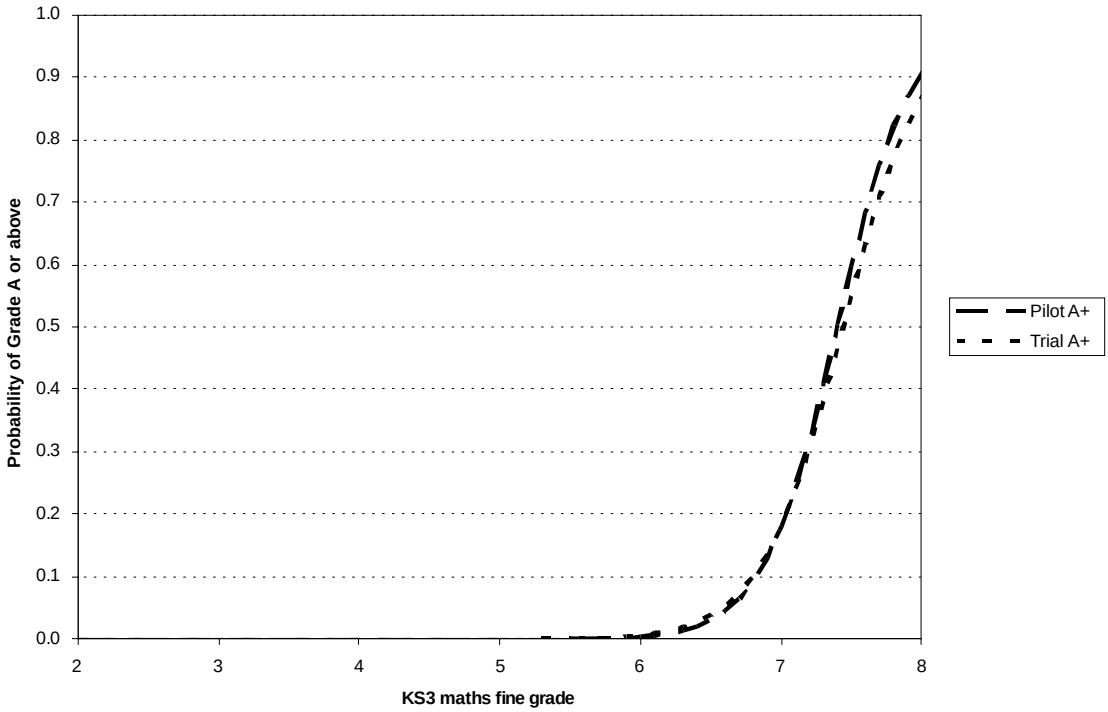
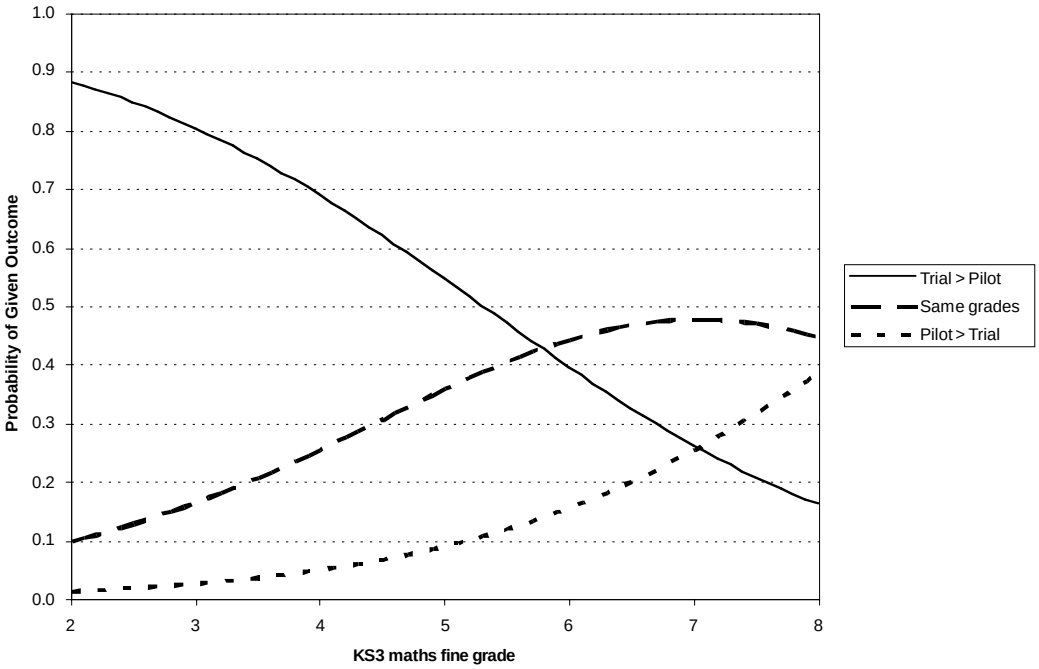


Figure 3d GCSE Grade A Probabilities as a Function of KS3 Maths Fine Grade



An alternative way of studying these relationships is to model the probability of obtaining a higher grade on the Trial or Pilot as a function of KS3 score. This is illustrated Figure 3e.

Figure 3e: Relative probabilities of achieving similar or different grades in relation to KS3 levels achieved



Thus, a KS3 candidates who obtained up to level 6 the Trial is advantageous. For students at level 7 the results are the same. The Pilot is only advantageous over the Trial for the very high scoring KS3 candidates, and even for these the probability of obtaining the same as the Pilot is higher than the probability of scoring higher on the Pilot.

3.3 What does the question level analysis reveal about the coherence of the papers and the actual difficulty of questions in relation to the intended difficulty?

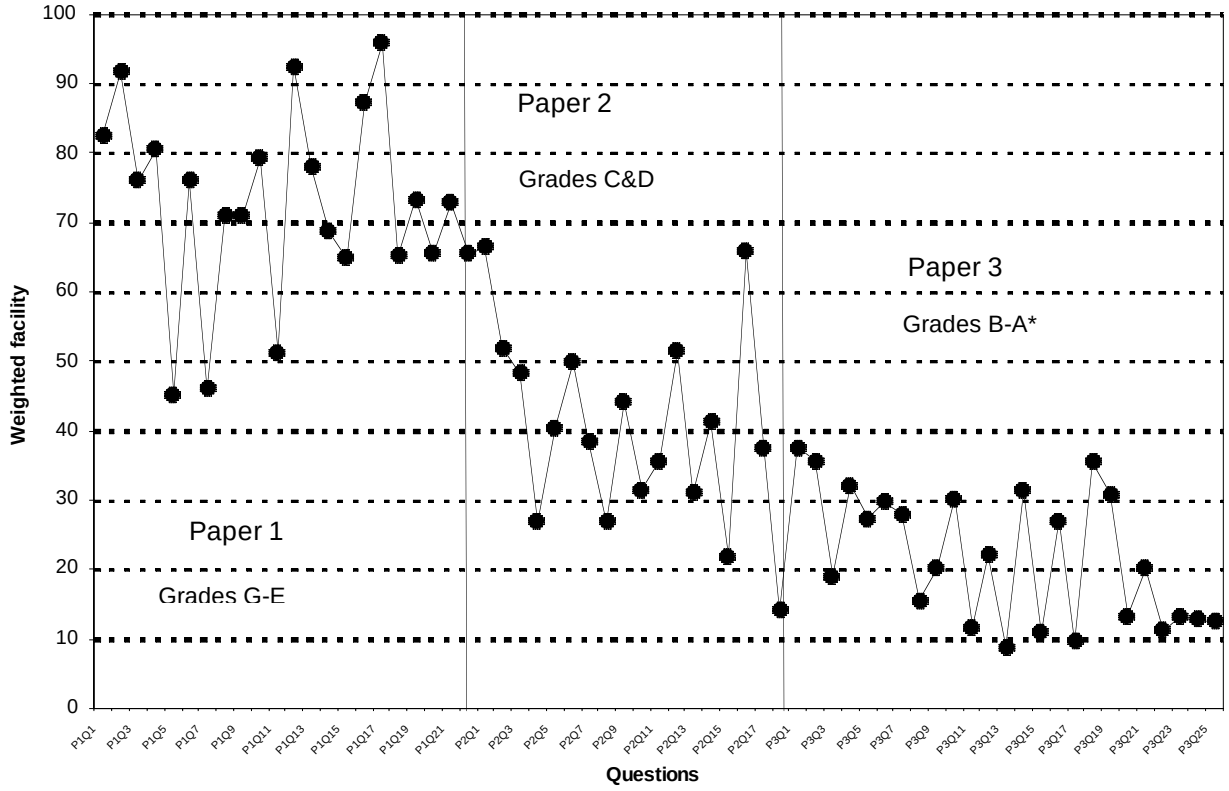
A series of analyses were performed to study the dimensionality of the items on each of the three papers in the Pilot. Simple factor analyses were carried out (see Appendix 5) and the variances explained by the first factor were 26%, 53% and 39%. This strongly suggests multidimensionality. The implications of this are that there is not a 'single maths ability' that will predict performance on different mathematical topics. We now study this from another perspective by looking at the relative difficulty orderings of the questions across the three papers.

One of the key assumptions of the Pilot model is that questions can be precisely targeted at particular grades in terms of both content and levels of difficulty. We explored this by analysing question level data. We also analysed Trial question level data (AQA) to see if this performed in a similar way.

While the question is straightforward, the methodology for answering is not. This is because not every candidate responds to every question, so the question difficulties have to be estimated on a common scale that takes account of the 'ability' of the candidates responding to each question. To estimate these difficulties (facilities) the 'pseudo-facility' of an item is the average mark obtained as a percentage of the maximum score on the item. Since these values depend on the ability of the candidates attempting each question a series of adjustments were made (see Appendix 6)

Figure 3f shows the weighted facilities, with items ordered in decreasing facility, i.e. increasing difficulty. It is clear from this that, although items tend to increase in difficulty across papers, this is not uniform and some Paper 3 items, for example, are easier than some Paper 2 items.

Figure 3f **QCA/OCR Pilot weighted question facilities**



Thus, for example, on paper 2 there were four questions that were harder than ten out of the 25 questions on paper 3.

For AQA a similar analysis was carried out for the Trial papers. Similar results for the factor analysis were obtained and also the same general results were found for the item difficulty orderings. However, for the Trial, the lack of a strict hierarchy is of less consequence than for the Pilot since the targeting of questions is more broadly based.

4. The awarding body meetings for the OCR Pilot and Trial examinations, July 2005

Because the essential difference between the Pilot and the Trial GCSE mathematics examinations is in how a grade is arrived at (see figure 4a), both the OCR awarding meetings were attended. The purpose was to investigate how the grades were determined with the different paper structures ('stepped' and 'overlapping'), what they have in common and what is different.

These meetings are reported in some detail in order to provide a clearer picture of how grades are set in GCSE examinations. This information is directly relevant to our analyses of what a grade means and of the reliability issues around grades.

Any comparisons between the OCR Pilot and Trial examinations are made simpler by some of the common features. Both sets of papers were set by the same Principal Examiners, the Chief Examiner was the same and the same examiners attended both awarding meetings, having marked scripts from both. These meetings had the same Chair of Examiners and the same Subject Officer in attendance. Both meetings followed the awarding procedures required by the QCA Code of Practice.

These similarities mean that the key differences were in the structure of the papers graded and the grading decisions that had to be made. These differences are laid out in Figure 4a (see also below for the differences in the target weightings of questions at each grade).

Figure 4a Summary of differences in Pilot and Trial structure and grading procedures

Pilot	Trial
Each paper is separately graded, combined with the coursework mark and the highest grade awarded	Each paper is graded, marks are standardised (UMS), and combined (inc coursework) to give the overall grade.
The grade range is a maximum of three grades on each paper	Each paper covers five grades
Awarding decisions at A,B, C,D, E, F boundaries (+ C/W)	Awarding decisions at A(2),C(4),D(2), F(2) boundaries (+ C/W)
Only one route to each grade	Two routes to grades C and D
Only 50% of the examination data used in determining final grade	All the examination data used in determining final grade

4.1 The Pilot award

The Pilot was first awarded in 2003. It is acknowledged by those involved that this first year was particularly problematical, partly because of the inclusion of Application of Number questions. This, plus the weighting of marks (A* 33%; A 33%; B33%), had led to depressed scores and grades (see Table 4.1). There was media attention also at the very low total marks need to secure a grade B (14%).While the 2004 examination was modified and results improved, the there was a perception that the award had still been too severe. The low grade B boundary (18%) was still seen as a public credibility problem – and raised reliability concerns.

In 2005 there were 7805 candidates, with 60 per cent taking the Foundation tier. The same schools were involved, so it was a similar cohort to the previous years.

Tables 4.1 (i-iii) Grade boundaries and cumulative percentages achieving grades in Pilot 2003-5

i) Grades A* - B (Paper 3)

Year	Grade A*		Grade A		Grade B		U	
	Mark*	% achieve	Mark*	% achieve	Mark*	% achieve		%
2003	54	6.7	37	13.5	14	68.2		31.8
2004	68	7.1	45	23.3	18	72.3		27.7
2005	79	7.7	56	27.9	24	73.5		26.5

- as percentage – paper out of total of 126

ii) Grades C - D (Paper 2)

Year	Grade C		Grade D		U
	Mark*	% achieve	Mark*	% achieve	
2003	35	58.8	21	75.4	24.6
2004	35	57.3	21	75.8	24.2
2005	39	57.7	23	75.3	24.7

iii) Grades E – G (Paper 3)

Year	Grade E		Grade F		Grade G		U
	Mark*	% achieve	Mark*	% achieve	Mark*	% achieve	
2003	56	66.3	40	87.2	24	97.3	2.7
2004	51	67.7	35	87.1	19	97.4	2.6
2005	60	67.3	46	87.3	32	96.2	3.8

Changes in 2005 – adjusted weightings

The 2005 awarding meeting began with the acknowledgement that this award 'has a history to live down', particularly in relation to the A*-B paper (Paper 3). One change for 2005 was that the equal weighting of questions (33/33/33) had been modified by QCA in April 2005 so that the target distribution of marks was **A* 20%; A 30% and B 50%**. On the grade C-D paper (Paper 2) the weighting had moved from 50/50 to ***C 40%; D 60%*.** Paper 1 had also moved from equal weightings to **E20%; F30%; G 50%** weightings. The reasoning for this was to bring it more into line with the Trial weightings. This had led to late changes to the examination papers.

Allowing more marks for the lower demand questions on all the papers had led to improvements in performance because the papers were more accessible due to the higher proportion of relatively low demand questions. The awarding dilemma this produces is how to interpret better performance on papers that give more opportunities to demonstrate what is known. The candidature was similar to the previous years, though there was a slightly higher proportion entered for the Higher tier – explained in terms of confidence returning after the 2003 difficulties. The assumption of the awarders was, therefore, that a similar pattern of results could be expected, though some allowance could be made for the relative severity of the previous years' grading.

The awarding process for the Pilot

The C-D paper (P2)

The meeting, in compliance with the QCA Code of Practice, began with the C/D boundary. This is the common paper which is taken by all candidates. It was recognised that this paper could be irrelevant to higher attaining students (A*/A) and dispiriting to candidates getting grades F or G since all the questions are pitched at grades C & D without any easier 'lead-in' questions. This was supported by the wide spread of marks (max.100) on this paper which had a mean of 47.6, a standard deviation of 28.0 and for which the most frequent mark (mode) was 20. Over 10 per cent of the candidates got 87 marks or more, while 25 per cent got 23 marks or less. This was an unusual 'flat' distribution of marks, best explained in terms of combining three distributions, the A*-B candidates having bunched high marks, the E-G candidates bunching at the lower end and the C-D candidates bunching in the middle – so that the mark distribution was almost evenly spread across the mark range.

The grade boundaries were set at 38% for grade C and 23% for grade D. The cumulative percentages of candidates getting these grades were 57.7% and 75.3%. A quarter of candidates were unclassified (U).

The A*-B paper (P3)

There was agreement among the examiners that P3 had 'got it right' in 2005 and had been much more accessible as a result of the change of weightings and the questions set. This was supported by the written comments of all 7 assistant examiners.

Examiners noted that candidates seemed to be getting marks throughout the paper, one examiner writing 'candidates seem to underperform in some of the easier areas of mathematics but then coped well with some higher level questions'. There also seemed to be some evidence of schools' selective preparation, for example an emphasis on trigonometry rather than algebra.

The mean mark on this paper (max. 126) was 52.1 (41.3%) with a wide spread of marks (SD 29.1). The mode was 32 with 25 per cent of the candidates scoring 23 marks (18.3%) or less.

The grade boundaries were set, after some statistical considerations, at 70 (55.6%) for A and 30 (23.8%) for B. Over a quarter of the entry received a U (Unclassified). The A* boundary is largely determined statistically (based on the A-B mark range) but in this case had to be adjusted to 100 (79.4%). 7.7 per cent of the P3 entry gained A*.

The E-G paper (P1)

It was noted that the 'bottom weighting' of this paper, with 50% of the marks targeted at G had produced a positive response with candidates attempting questions throughout the paper.

The mean mark on this paper (max.100) was 64.8, with most of the marks bunched around this (SD 16.1). The mode was 69 and 83% of the candidates scored over 50 marks. The grade boundaries were set at 60 for grade E (67.3% of the entry for P1); 46 for grade G (87.3%). The arithmetically determined G boundary was 32 (96.2%).

Comments

The Pilot award informs several of the central questions of the evaluation. A key question is *what does a grade mean?* Issues here are what can be inferred from the award of a grade and how reliable is that grade.

On the Pilot, the grade is the best result from the two separate papers. The grade is therefore the result of performance on questions intended to make demands appropriate for a relatively narrow band of grades (A* -B etc). While the analysis of the sample data and results of the QCA script scrutiny (not reported here) provide more detailed findings, it was apparent from the review of scripts that candidates, particularly those who score relatively few marks, rarely gain their marks in a predictable fashion. For example a grade D candidate may not have successfully answered the 'typical' D questions but have got marks on the more difficult C questions. Similarly a grade B candidate is unlikely to have gained all the marks on 'grade B' questions, some were gaining part-marks on some of the most difficult questions while missing questions designed to be easier.

The relative low grade A boundary (55.6%) suggests a grade A candidate has either not collected all the 'grade B' marks (50% of total) and has gained 'grade A' marks or has done well on the B questions and got just a few marks on the grade A questions. Similarly it was technically possible to get an A* (79%) by getting all the grade A & B questions correct (80%).While this is a more acute problem for the Trial papers, it is not negligible for the Pilot.

A related issue is the impact 'width' of the grade boundaries on the inferences that can be made about the mathematical competence of a candidate. For example a candidate who got 54% on Paper 3 would have got a grade B, as would one who got 25%. Both would be treated as equal in that they got the same grade, but there is likely to be considerable difference in attainments.

A concern linked to this is the reliability of a grade based on only 50% of the total marks (i.e. a single paper), especially when only a quarter of the marks are needed on this one paper (grade B, grade D) and these may have been gleaned from part-scores on many questions.

The second theme is *how positive the experience of the mathematics examination was for the candidates.* A structural dilemma for the Pilot is that the C-D paper is likely to be seen as a waste of time for A*-A candidates and too difficult for grade F and G candidates. The examiners' assumption was that over time some students would only enter one paper (as in Scotland). The Pilot status of the examination had required all students to enter two papers. The relative difficulty of P3 for grade B candidates and the impact of a quarter of the candidates failing both the P3 and P2 are likely to have some repercussions on attitudes towards mathematics. Our student survey data (see Section 6) provides more insight on this.

4.2 The Trial award (OCR)

The Awarding Panel involved the same personnel as the Pilot panel. This provides a mechanism for ensuring comparability of demand between the two examinations. Because it was a new examination for which there were no previous standards on which to base grading judgments, it was made clear from the outset by the Chair of Examiners that the task was to produce similar overall grade distributions to those the Trial candidates had achieved on their other GCSE mathematics examination – 3-tier or Pilot. These were available and were used

as the basis of modeling the consequences of the proposed Trial grade boundaries. There was some discussion of whether this strict equivalence was appropriate, given that some Foundation tier candidates may have achieved grade C on the Trial, a grade not available on the 3-tier examination, therefore a higher percentage of grade Cs might be anticipated. This was ruled out.

The Trial papers had a more general approach to the targeting of questions. On the Foundation papers 75% of the marks were 'low tariff' and intended for performance at the grade F – G range. 25% of the marks were 'medium tariff' and were based on grade C- D work. These questions were also included on the Higher tier paper in order to offer some comparison in setting the C grades on both tiers. The Higher tier papers were largely targeted at grades D and C (55%) with 25% allocated for grade B, 12% for grade A and 8% for A*.

The Chief Examiner asserted at the outset of the meeting that, while these weightings facilitated awarding at grade C, it made the award of grade A difficult because of the more limited data. He considered that this tier 'would need to be made harder' in future as at the moment it was too easy for the most able students; the same point was made by teachers (see below). It was noted that over half the candidates for the Trial were entered for the Higher tier (this is different from the other Awarding Bodies for whom the majority were Foundation tier). An analysis of their grades on their other GCSE mathematics papers (3 tier or Pilot) suggested the candidature was skewed towards the higher attaining end of the grade distributions (see Table 4.2).

Table 4.2 Cumulative grades achieved by Trial candidates on their other GCSE Mathematics examination and on Trial (Awarding meeting distribution*).

Exam	A*	A	B	C	D	E	F	G
Other	4.9	18.1	38.1	63.5	77.9	90.1	94.6	96.4
Trial	4.5	18.4	38.0	63.8	77.8	88.3	95.0	98.7

***** This was based on data available at the meeting, when all the results were incorporated the Trial percentages were lower for the higher grades (see Table 4.1)

The awarding process for the Trial

In line with the QCA Code of Practice grading began with the grade C boundary on the Foundation tier. Decisions were made on each paper, the coursework boundaries having been predetermined. The percentage of candidates at this boundary on each paper was kept broadly equivalent (by adjusting the grade boundaries). The grade C boundaries on the Higher tier papers followed and this was acknowledged to be more difficult for awarders as there was less information and some grade C candidates gained marks on high tariff questions. The grade final boundaries (max 100) were: Paper 1 (calculator), 69; Paper 2, 72; Paper 3 (Higher tier, calculator), 31; Paper 4, 34. These boundaries were each 3 marks lower than those arrived a judgmentally by the examiners.

This was followed by the grade A boundary. The limited material at this level (20%) was seen as problematic, especially as the high tariff probability question had been reasonably well answered by grade B candidates. It was noted that many grade A candidates dropped marks on the easier material (e.g. a stem and leaf question on which other candidates often

did well). The mean mark of 62.4 on Paper 4 (non-calculator) was considerably higher than Paper 3 (55.2) and this was put down to some of the more accessible topics being on this paper (there had been a similar gap on the three-tier papers). The A boundary was, after a number of statistical modeling iterations, put at 67 for Paper 3 and 77 for Paper 4, each two marks lower than the judgmental recommendations. The A* grade boundary is statistically determined and this was driven by the need to equate the percentage gaining A* with that of the other GCSE taken (4.9%). This was therefore fixed at 81 (P3) and 89 (P4) –with examiner concerns that this could be achieved without any marks from the one A* question.

The process was repeated for the F boundary, though the agreed mark of 42 on Paper 1 was higher than had been proposed on the basis of professional judgement (35). This pattern was repeated for Paper 2 (49).

The agreed grade boundaries are set out in Table 4.3

Table 4.3 Trial grade boundaries (max. 100 marks)

Paper	A*	A	B	C	D	E	F	G
1				69	60	51	42	33
2				72	64	56	49	42
3	81	67	49	31	21			
4	89	77	55	34	18			

Comments

The papers were generally thought to have performed well, with the grade boundaries in, or relatively near, the target thresholds (e.g. A 75-65; C 45-35 on higher tier). All the examiners at the awarding meeting were exercised by the relative lack of questions at grades A and A* levels of demand. This meant that, in theory, a candidate could get a grade A without doing any grade A questions. This is compounded by them often dropping marks on other questions so the A boundary allows for ‘dropping’ 25 % marks.

In relation to ‘what does a grade mean?’ this means that only restricted inferences can be drawn. It can be said with some confidence, because of the wide range of questions across two papers, that the grade A candidates’ general mathematical attainments place them in the top 20% of the group examined. What cannot be inferred with any confidence are the particular mathematical competencies that have been demonstrated.

5. Analysis of the teacher questionnaires and interviews

In this section we explore the data from teachers. We are aware that interest is in the relative merits of the two 2-tier models. Only those teachers from the 8 schools involved in both the OCR Pilot and the OCR Trials were able to directly compare the two models (see section 6). All other teachers are commenting on the relative merits of 2-tier Trial in relation to the 3-tier papers.

We comment briefly on the ways in which the Trial was administered before reporting on teacher and student perceptions of the Trial examination.

5.1 The administration of the Pilot & Trial: patterns of entry

For the Pilot: For students in schools taking part in the QCA/OCR Pilot this was not an issue – with very few exceptions they took the same tier in both cases.

For the Trial: In nearly all cases 'Foundation' students were entered for Foundation papers and 'Higher' students were entered for Higher papers (one '3-tier Higher' candidate took a foundation level Trial exam and four '3-tier Foundation' candidates took a higher tier Trial exam). Practices for entering 'Intermediate' students varied (see Table 5.1). Because this was a one-off examination about which decisions had to be taken quickly we do not see these as necessarily stable patterns.

Ways of deciding how to enter 'Intermediate' students

Most schools used some combination of predicted (or target) grades with students predicted B or better automatically being entered for the Higher tier and 'C/D students' being entered for Foundation. (although these grades were also available through the 'Higher' route).

Table 5.1. Placement of 'Intermediate' candidates in trial examinations

Board	3-tier Intermediate	2-tier Higher		2-tier Foundation		Number of schools entering all Intermediate candidates for the same examination	
						Foundation	Higher
AQA	374	193	55.6%	154	44.4%	1	0
Edexcel	213	84	42.2%	115	57.8%	0	1
OCR	121	55	46.6%	63	53.4%	0	0
WJEC	66	34	58.6%	24	41.4%	0	0
Total	**774**	**366**	**50.6%**	**356**	**49.3%**	1	1

Data missing for some records so the total of the two, 2-tiers ≠ number of Intermediate 3 tier candidates

Of those schools commenting specifically on students predicted a C grade, half entered them for the Higher tier on the chance they could get a better grade, the other half entered them for Foundation so they would be 'sure of getting a C'. This latter decision seems very conservative and one teacher expressed regret in retrospect – it would also indicate something about teacher expectations.

A few schools left the decision about whether to try Higher or Foundation papers to their students – generally with some discussion. They appear to have made a variety of decisions – either to have a go at a higher grade, or to have a go for a C on a paper that was anticipated to be easier.

Perceptions of degrees of difficulty

We asked teachers to comment on their students' responses to the examinations. The extent to which they were reporting student perceptions or their own perceptions are not clear. We summarise these responses briefly (Table 5.2):

Table 5.2 Teachers' perceptions of relative difficulties of examinations

3-tier									
Board	Higher			Intermediate			Foundation		
	easy	about right	hard	easy	about right	hard	easy	about right	hard
AQA	1	18	4	2	19	1	1	16	0
Edexcel	2	18	3	2	21	4	2	17	1
OCR	0	8	3	0	13	0	0	6	0
WJEC	0	10	1	0	7	2	0	7	1
Total	***3***	***54***	***11***	***4***	***60***	***7***	***3***	***46***	***2***
%	***4%***	***79%***	***16%***	***6%***	***85%***	***10%***	***6%***	***90%***	***4%***
Trial									
Board	Higher						Foundation		
	easy	about right	hard				easy	about right	hard
AQA	6	17	7				5	14	0
Edexcel	10	17	0				12	16	0
OCR	8	7	0				1	6	0
WJEC	6	6	0				0	8	0
Total	***30***	***47***	***7***				***18***	***44***	***0***
%	***36%***	***56%***	***8%***				***29%***	***71%***	***0%***

Note that some teachers commented for papers 1 & 2 – this results in some double counting.
Percentages refer to the comments on that tier.

There was a general perception that both Trial examination papers were too easy. This might be expected – many candidates sitting the Higher Trial papers were given easier questions that they were used to (from Grades C and D), while any 'Intermediate' candidates entered for the Foundation papers would also see much easier questions than they were used to.

Approximately one third of the comments related to *the level of difficulty of the papers.* Mostly this translated into the Higher papers being too short and/or too easy for 'Higher tier' students. This was counterbalanced by some who felt the paper was better for 'Intermediate tier' students – or would be better if they had been taught more of the 'Higher' curriculum. Similarly the Foundation paper was 'too easy/short' for 'Intermediate' students but provided a good challenge for 'weaker' students.

Some commented that *one paper was significantly easier/more difficult than the other* – that the papers were unbalanced in some way. Indeed a large number of the comments (more than 50%) included the observation that *all* the papers were either too short or the time allocation (2 hours each) was too long. Reports of the time taken to finish indicate that most students

finished the papers in between 40 minutes and an hour and that the longest anyone needed was an hour and a half.

There were also *positive comments* – that the questions were mathematically sound with a good selection/variety of questions. Teachers and students also liked the 'lead in' that the broader range of questions provided, the easier questions at the start built students' confidence. (There were one or two who complained that more able students were 'thrown' by this as they tried to over-complicate the questions but this would clearly not happen if the students were prepared for the papers)

5.2 Teacher perceptions of the experiences of students studying mathematics at GCSE

The Smith enquiry raised concerns about students' experiences of learning mathematics. These in turn reflect tensions inherent in the dual purpose of the GCSE: as certification for the cohort completing compulsory education, and as a selection tool for GCE A level mathematics. The results of the teacher survey indicate that teachers are similarly exercised about these issues and the way they play out in relation to the examination structure.

Teacher concerns and perceptions about student experiences.

Virtually all teachers (and students) commented that the two hours allowed for the papers was felt to be far too long (all papers, all exam boards, both tiers). A smaller number suggested that the time allocated to the Foundation papers to too long, whereas for the Higher tier papers there were not enough questions. This may be regarded as a teething problem which can be addressed rather than a structural one.[1]

Certification and the experience of lower attaining students

In agreement with Smith, most teachers and their students find it wholly unacceptable that students entered for Foundation level cannot currently achieve a C grade however hard they work or however well they do. As one teacher said:

> "The current 3-tier system demotivates students by denying them the opportunity to better themselves. Students entered for the Foundation tier are more likely to switch off once they know the best grade they can get is a D. The reality of it is that we try to keep the options open for as long as possible but eventually a decision has to be made and once entered for the Foundation tier we have lost some of them. This has had a knock on effect for the others. The 2-tier system will hopefully resolve these issues offering them an opportunity to get the C even if the chances of them getting the grade are somewhat remote. The psychological factor is not to be underestimated."

Most keenly anticipate the availability of a C grade in the Foundation and the motivation this will provide for the many who currently 'switch off' once they realise they can only get a

[1] Though there is anecdotal evidence that examiners feel constrained by QCA's *Target & Tariffs* document which allocates specific marks to specific topics, so that it may only be possible to ask one question on a particular topic because few marks are allocated. There is also a perception that only a fixed number of questions can be asked in association with particular grades and that the time per grade available was fixed – so a two hour examination was believed to be mandated for a paper covering four grades.

'D' grade. They anticipate better teaching of this cohort of students under two-tier (Trial) conditions – since these are the grades that are important for reporting and accountability there would seem to be strong, positive implications for 'Foundation students' and their teachers.

Comments about the two-tier trial were overwhelmingly positive:

> "Great that everyone has the chance of a C – much better motivation for lower ability students, parents will be happier too"
>
> "The lower tier will help motivate students if they see a grade C is possible, this should result in a more positive learning environment in the classroom."

There were very few negative comments. Those there were related to the experience of taking a 'harder exam', particularly for those predicted and E or F grade.

> "The Foundation paper will cover more challenging topics which will make the exam more difficult for weaker students"

The experience of higher attaining students

Again echoing Smith, while teachers were overwhelmingly positive about the move to a two-tier examination structure for the less able they were notably more cautious about those they felt to be more able. Concern for the 'more able' was sometimes expressed directly, at other times it was expressed more indirectly in relation to A level.

Teachers were worried that the Trial examination papers offered nothing to stretch or challenge the most able, or to distinguish the A* candidate. However, this is inevitable if the Higher syllabus is changed or made more accessible to more students through covering a wider grade range.

> "Doesn't seem to stretch the A* top students as there are less questions aimed at this level. My students did not enjoy covering some of the D-grade topics e.g. rotational symmetry when they'd spent 2 years doing things like linear and quadratic simultaneous equations."

This carried on into concerns that these students wouldn't be as well prepared going into A Level. There were some calls for the old extension papers to be reinstated – or anticipation that they would be entering most able students early and starting them on AS in Y11. (More on this below)

5.3 Teacher and school level concerns.

How will I know if a student will be able to cope with the A level syllabus?

Smith (2004, para 4.10 and recommendation 4.6) reports concerns over the ambiguity associated with the two routes to gaining a grade B: via both the Intermediate and Higher Tier. He notes that by studying the Intermediate Tier, a student can achieve a B grade through a course of study that includes "*significantly less … algebraic and geometric content*" than a peer studying for the Higher Tier. This difference has particular importance as a B grade is generally accepted as a necessary minimum requirement for AS/A-level mathematics.

In relation to the current Trial, there seems to some variation in the GCSE grade deemed to provide adequate access (or to sufficiently demonstrate ability) to an A level course. The range of practices include:

> Only accepting A or A* candidates onto AS/A level courses (Trial or Pilot will make no difference to this)
>
> Accepting students with a GCSE grade B on to AS courses – some with the proviso that the 'B' comes from Higher tier papers. (A potential issue for the Trial model)

Those who expressed concern about this were worried that it would be easier to obtain a B on the 2-tier papers and that many would then struggle on AS, this might in turn lead to a higher drop-out rate. A similar concern was that if it was 'easier' to get higher grades on the 2-tier examination then students might who appear to be able to do enough maths to cope with A Level would struggle. This is discussed in more detail below (S 4.4.1). The Pilot model might be expected to avoid this but it can result in a narrower curriculum being taught so students may not have the breadth of understanding (See section YYY). Teachers also raised this issue in relation to the possibility of an extension programme (S 4.5.3).

Tactical behaviour in setting and exam-tier entry

Smith draws attention to, and laments, the *tactical behaviour of schools and students* who believe that it is easier to get a B grade on the Intermediate than the Higher Tier.

> *We have been informed that when grade B was first introduced as a possible outcome on the Intermediate Tier, entries for the Higher Tier fell from nearly 30 per cent to about 15 per cent of the candidate cohort and have remained relatively stable since then.* (para 4.11)

This tactical behaviour is understandable if read in relation to the importance of school league tables – if it is easier to get B grades from the Intermediate tier then, while the measure of success is A*-C grades it makes sense to enter students on 'safer' papers. However, the implications for teacher expectations and students' futures may be problematic. Some schools visited reported ex Intermediate students who had achieved a 'B' grade being unable to go on to study AS mathematics because the local 6th Form College would only allow this if the Higher tier had been studied.

In terms of the move to the two-tier examination structure, the issue is recast – there will be only one route to a B grade, but the equivalence of the C will be called into question. The single route to a grade offered by the Pilot model would appear to get over this problem but teachers from Pilot schools visited report engaging in highly targeted teaching. This approach is based on an assumption that topics 'belong' to particular grades – generally that anything 'algebraic' belongs with 'Higher' tiers while 'arithmetic' is a 'Foundation' concern, specifically (for example) Pythagoras is a 'grade B' question, while stem and leaf diagrams are 'grade D' questions. This makes life much easier for teachers as they feel able to teach a greatly restricted curriculum, but it raises questions about the impoverished curriculum being experienced by the students.

Teachers also anticipate that a degree of tactical teaching may become necessary if the Trial model is adopted but it is of a different kind, drawing from a wider curriculum base. (It is also possible that these are rationalisations of teachers anticipating a change that they have not yet worked with). The concerns relate to setting arrangements :

> "Where do you stop teaching students who are not Foundation or Higher but have a target grade of a 'B'? Some of the Higher content is going to be too difficult for them – they could achieve a 'C' on Foundation but be below their target."
>
> "Concern that grade 'B' students will opt for Foundation paper as it is easier – and the higher paper may appear inaccessible because of the harder topics: trig equations, transformations of graphs, vectors etc."

The concerns also relate to curriculum content: the 'overlapping' Trial model requires teachers to take a broader approach than the Pilot model and while they may decide not to teach the whole syllabus to all groups, there would appear to be more flexibility anticipated

> "some groups might only pay lip-service to certain grade material"
>
> "We would no longer have a tier called 'intermediate' although we would probably still have the type of set because we wouldn't teach topics at the higher end. We could probably teach to B rather than A for the lower sets – but not sure yet. Not sure how parents would react."

Similar arguments were made about F and G candidates faced with C and D material in the new Foundation tier:

> "We may become more selective of topics taught at higher tier to C or B grade candidates and at Foundation tier to E or G grade candidates."

Another few anticipated "more teaching of 'the basics' and numeracy and less emphasis on the higher grades." Which may be a good thing, or may reduce the challenge.

5.4 What effect will the move to two-tier examination structure have on those who might potentially go on to study at A-level?

> "*If there is no significant restoration of the numbers entering AS and A2 mathematics within the next two or three years, the Inquiry believes the implications for the supply of post–16 qualified mathematics students in England, Wales and Northern Ireland to be so serious that consideration should be given by the DfES and the relevant devolved authorities to offering incentives for students to follow these courses.*" (Smith, 2004, Recommendation 4.8)

The concern about take up at A level relates wholly to the 'selection' function of the GCSE examination and draws on concerns about the standard of the new two-tier examination.

Effect on uptake at A/S and A Level

While recognising concerns about the impact on take-up rates at A level (Table 5.3), there were also many positive reactions. There were some who commented that their current strong 'B' grade Intermediate candidates were currently barred from AS maths although they would like to take it – this barrier will evaporate with a 2-tier examination structure. A few expressed unqualified pleasure at anticipated increased numbers and its effect on student morale and motivation:

> "Under the 3-tier system the higher students are achieving low marks during the course and final exam but are still achieving B or C grades. During this period their confidence is being drained away as they continue to fail"

"Increase rates of take-up since students would feel more positive about maths at post 16"

Table 5.3. Teachers' feelings about the effect of a move to a two-tier examination structure on numbers going on to study A level

	Trial	Pilot	
Don't know	6	0	
None	10	0	Comments here generally suggest schools expect students to get a B before they'll take them onto A level courses and they don't expect the numbers achieving this to change although some anticipate they may be better prepared if they have taken the 2-tier exam – presumably because if there is only be one route to a 'B' everyone will have done enough algebra.
Increase	35	4	The belief that numbers will rise came with warnings from nearly all respondents that more students might feel inclined to enter due to the greater availability of higher grades. However, teachers suggest that many will struggle as AS & A2 are significantly more difficult – particularly in relation to the algebra. There was some feeling that the 2-tier structure will make people 'over' confident and/or that it will not provide such a good preparation/ grounding as the 3-tier Higher paper. There were a small number of teachers saying that more accessible papers will enable more students to have positive experiences of maths which may lead to more wanting to continue
Little	17	0	Most report this as a small increase and cite reasons similar to those above. A very few commented that the most able – those who will go on to succeed at A level – will be unaffected by the changes as they would have taken the Higher papers anyway. One or two believe that dropout rates may rise as students are lulled into believing they are better at maths than they really are – this is founded on an assumption that A grades will be easier to attain.
Lower	2	2	This was generally unexplained although some schools appear to anticipate entering all students likely to achieve below a B grade into Foundation levels leading to fewer students doing Higher tiers. (from a 2&2 school)
N/A	2	1	This from schools that don't teach A level (including special schools)
Unclear	3	1	

Those very few who anticipate numbers falling were concerned about current potential 'B' grade students being limited to (or opting for) a 'C' at Foundation as a safer route than the chance of a B (or even a C) at Higher:

"Grade 'C' Foundation not good enough preparation for AS/A2. Perhaps our numbers will drop if 'B' students opt for Foundation!"

"Those who achieved 'B's at Intermediate level might be restricted to 'C' as safe route and not go on to A level".

That this is voiced at all is worrying – if teachers choose to place increasing numbers of 'Intermediate' students into the Foundation tier because they believe it will be easier to get a 'safe C' this way will severely limit students' learning. These kinds of tactical behaviour are a potential problem with both the Pilot and Trial models.

If it becomes easier to get an A*, A or B on the two-tier papers then there is a worry that less able students will believe they are able to go on to A level and then struggle:

> "I am concerned that this [move to 2-tier] will weaken the standard of mathematics expected to be covered at KS4 and the knock-on effect on abilities at A Level."
>
> "The coverage of the curriculum at the higher level may be less rigorous. Thus, students who chose to study A level may not have the necessary algebraic skills demanded for success at A level."
>
> "The exam cannot be watered down again. It has such a knock-on effect with AS, A and university level."
>
> "If the higher level is that on these H papers the preparation for AS will not be good enough."

There were particular concerns about the amount and type of algebra that students study at GCSE as this is felt to be the chief difficulty at AS/A level.

There was a lot of concern that the 2-tier Trial exam structure would prove to lack challenge for the most able students.

> "I am concerned about the A/A* students. We may have to enter them for A level modules to keep them challenged."

Certainly the 'stepped' Pilot structure allows teachers and the most able students to treat the C/D paper as an insurance policy and to focus on the A*-B curriculum. However, as noted above, this can severely restrict the curriculum that children experience. Most (Trial) schools felt there were ways in which they could manage this potential loss of challenge (see below).

5.5 The depth and breadth of the syllabus and the place of coursework.

There is a recommendation (Smith, 2004: 4.3) that the quantity of coursework in GCSE mathematics be reviewed "*and, in particular, the data handling component, with a view to reducing the amount of time spent on this specific element of the course.*" And (R4.4) that the content of the curriculum should be reviewed (in relation to data handling & statistics – which remain important) so that more time can be given back to "*the reinforcement of core skills, such as fluency in algebra and reasoning about geometrical properties*".
Concerns reported by teachers relate to coursework, the content of the curriculum, and providing challenge for the most able.

Coursework

Many teachers expressed concern over coursework (this was a spontaneous expression – there were no prompts about this) – it is certainly an unloved aspect of the GCSE. Several reported that the availability of the internet made it very easy for those with access to find exemplar answers and that this, in turn, renders this element of the examination largely meaningless.

> "It'd be even better if course work was abandoned."
>
> "Eliminate coursework – it doesn't add to the educational experiences of student or of teacher."
>
> "Coursework in no way reflects the 'mathematical' ability of students."

There were concerns that, because 2-tier papers covered a wider grade spread and so there were fewer questions targeted at particular grades, the papers will be less able to give 'real' indications of ability raising the importance of coursework. This was dreaded:

> "These papers do not differentiate the students. Coursework will have a bigger impact on grades – frightening."
>
> "With coursework, which is not the best discriminator (and must eventually go, surely), there may not be enough discrimination for the very best."

A suggestion for testing students' ability to apply their knowledge made by several teachers would be the inclusion of a timed investigative task sat under examination conditions.

Curriculum content

There was a general feeling from the questionnaires and the interviews done with teachers involved with the Trial that the move to a two-tier examination structure would give students better access to more of the mathematics curriculum. Some expressed concern that this might lead to curriculum overload and that it may need to be rethought to some extent:

> "I hope the amount of algebra Intermediate (→ Foundation) students are expected to learn is reduced and more emphasis placed on numerical skills, which will be of more benefit to many."
>
> "'Higher' students will get taught more basic maths/numeracy (still important)"
>
> "More curriculum to cover – and where/whether to make cut-offs for ex-Intermediate/ Foundation students who will now be confronted with more challenging work?"
>
> "Students will get the chance to do a course that is actually designed to meet their needs so students will see the point of learning the subject and will engage."
>
> "Should the curriculum be reduced or changed in some way?"

There were only two comments spontaneously made about statistics in GCSE and the nature of the concern appeared balanced:

> "The inclusion of statistics has been reduced in the 3-tier system due to coursework requirements. I am concerned this change may reduce statistics even further."
>
> "The statistics coursework is a statistics project not one suitable for the general mathematics student."

Maintaining challenge for the most able

Clearly, the move to a Trial two-tier examination will involve each 'tier' and each examination covering more of the curriculum. If the main purpose of the examination is certification the bulk of the questions must match the majority of the students taking it.

Smith (Recommendations 4.5, 4.8 and 4.10) suggests that schools should review measures to "*support and encourage current GCE course provision for the most able mathematics students"* and that an "*extension curriculum and assessment framework*" will need to be developed. It has been suggested that the Pilot examination structure answers this call with C/D paper acting as the core and the A*-B paper as the extension. Given the current examination's status as the terminal examination for most 16 year olds, such a reading is

problematic. If the extension curriculum and assessment framework is part of the statutory curriculum the limited access to this would raise equity issues – back to the tension between the certification and sorting functions of the GCSE (some teachers had suggestions about ways around this – see below).

Several schools had clearly thought about the issue of lower challenge for the most able resulting from examinations covering wider grade range. Some asked about the return of specific 'Extension papers' to be provided by the exam boards. Others felt able to meet the challenge within the existing framework:

> "If coursework could be axed we could provide our own extension/exploration work and probably enter some sets in Y10"
>
> "I would expect more schools to fast-track top set and start AS work early."
>
> "[there will be] dumbing down at the top end – but we can provide extension work."

If an extension paper is provided by the exam boards it may become expected that all students take this – it would need to be beyond the scope of the reporting and accountability structures to prevent 'grade inflation' and over testing. The option of taking GCSE early and starting AS early would appear more attractive – it does not increase the number of exams available. Anyone planning to do AS maths (only) as a support for A levels in, for example, geography, economics, or biology has the option to either get it done early or to take two years over it.

5.6 Resource implications

Teachers who responded to the questionnaire reported no concerns about the proposed change to a two-tier model. Teachers who have been using the OCR Pilot since 2003 reported no technical difficulties with implementing the change.

About 90% of respondents believe the move to a 2-tier exam will be quite easy or very easy.

Very few resourcing issues are anticipated. Rather more than half think they will need to buy new text books. Most of the rest think that what they have will be fine – the few others don't know yet.

Many were anticipating the extra work that would be involved in rewriting the department's Schemes of Work (SoW) and the extra meetings that would be needed to plan the tiering. Whilst this appeared irksome for many, it was not that they did not know how to do this work – rather the opposite.

Few mentioned professional development (CPD) needs – help re-doing SoW was one, meetings to gain familiarity with new papers and expectations was another area that schools might appreciate some support with.

The general feeling was that the change will fit well with existing arrangements and school structures being more in line with other curriculum areas and cause no extra work for examination officers (indeed, the Trial model will simplify matters for schools as there will be fewer different papers to organise). It was also felt by many that the similarity with other examination structures will make it easier for parents to understand. There will also be a better fit with College requirements.

5.7 The effects of the proposed examination structure on teaching and learning programmes.

The overwhelming majority of schools report setting students for mathematics on the basis of ability (the exceptions were a special school for students with EBD and the two FE colleges) – they anticipated no real change in this practice although the group labeled as 'Intermediate' would be lost. Many schools anticipate benefits from greater flexibility over which set to place students in; many pointed out that although there will only be 2-tiers there are likely to be more than two Schemes of Work. While movement from Foundation to Higher would be problematic because of the different curriculum, movement from Higher to Foundation would not be difficult. A few anticipating the move to the Trial model were looking forward to the possibility of more 'mixed ability' teaching – our sense is that, by this they mean teaching a wider range of abilities within one class ("more mixed ability teaching within broader tiers"). This runs counter to the developing trend or a narrowing curriculum identified in Pilot schools.

For some teachers what to do with 'B' students was anticipated to be the difficult decision:

> "It feels like it'll be harder to place 'B' students – do we put them in for Foundation for safety or Higher and let them struggle with A and A* work?"

For others, students predicted a 'C' became the issue with some reporting increased worry that it would be even more important to make sure students were entered for the right exam– this raised the question of whether a C would be the same, or equally easy (or difficult) to get on the two tiers.

Some schools reported that they will be more exercised over the placement of borderline B/C students who it was felt would struggle with higher level topics.

> "There are bound to be problems with borderline candidates between the 2 tiers. Should they be entered for higher or Foundation? Some students might find they can't answer a lot of questions on the papers and be discouraged by failure. It is really a question of swings and roundabouts. There will be gains and losses. Hopefully the gains will outweigh the losses."

6. Teachers' and students': a direct comparison of the Trial and Pilot models

This section relates only to responses from schools in which students sat both the Pilot papers and the OCR Trial papers. This is the third year that they have used the Pilot examination (and their only experience of the Trial model). For some relatively inexperienced teachers this meant that they had only taught the Pilot model.

6.1 Which of the two-tier models did the teachers who experienced both prefer?

Data from the questionnaires and interviews indicate a high level of ambivalence about the two forms of the two-tier examination and no clear, single message emerges (Table 6.1):

Table 6.1: Summary of the perceived advantages and disadvantages of the two models of overlapping two-tier examinations

Pilot	
Advantages	**Disadvantages**
A and A* candidates can 'bank a C' on a very straightforward paper and focus their learning on the higher syllabus then enjoy a more challenging paper.	The experience of most students taking this series of papers is not positive. A student predicted to receive a grade B or a borderline B/C candidate will experience 2/3 of the 'higher' paper as inaccessible (similarly a student predicted E, F or G will find the C/D paper very difficult).
The most able can be well prepared for A-level study	
Particular topics become associated with particular grade levels – this enables teachers to target their teaching. (One school we visited seemed very deliberate and proficient in doing this)	Some doubt it is possible to target questions as specifically as this format suggests/ requires
Highly targeted teaching can in turn reduce the weight of the syllabus to be covered	Highly targeted teaching can lead to a narrowing of the curriculum experienced by groups of students
There is a unique route to every grade	Because the marks are not aggregated, one paper is effectively 'lost' or 'wasted'.
Major issue raised: what is the purpose of the GCSE? If it is to act as the exit exam from compulsory schooling for the majority of students aged 16, then a series of papers that are experienced so negatively and which seem only to be of benefit to the top (A*=4%, A=12% 2004) would appear unacceptable.	

Table 6.1 ctd.

Trial	
Advantages	**Disadvantages**
'Setting' can be potentially less rigid within schools	There is no unique route to a C grade raising the question 'what is a C?'
Each paper contains a wide range of questions providing a 'lead in' and boosting confidence to 'have a go' at more challenging questions	Each paper contains a wide range of questions so the percentage of questions targeted at particular levels is reduced (does the Higher paper ***really test*** an A* candidate?)
More students have a positive experience of the examination.	Potential 'A and A* students' may be less challenged by the examination
More students are exposed to the 'higher' syllabus	Potential A level students may be less well prepared
Major issue raised: what is a grade C? This is the same as the 'what is a grade B' issue raised by the overlap between Higher and Intermediate tiers in the 3-tier system. However, as the C is not critical for A level uptake this would seem perhaps less important. The strategic entry of students into the 'new' Foundation tier if it is perceived that it is easier to get a C through this route could lead to fewer students taking Higher tier maths.	

6.2 Students taking two 2-tier examinations: QCA/OCR Pilot and Trial

430 student questionnaires were returned from schools which have been using the QCA/OCR Pilot examination structure since 2003. None of these students have any sense of themselves as 'Intermediate' since they have not been part of a three-tier examination structure. They are comparing the two, 2-tier models being evaluated.

In interview these students had little to say about the two models – perhaps, as a C grade is available to all and no one is loosing their mathematical 'home', they had difficulty relating to the difference between the exams being an issue.

There was a slight increase in the proportion of students sitting the Higher tier in the Trial (Pilot: 48% Higher, 52% Foundation; Trial: 51% Higher, 49% Foundation). This is likely to have been borderline C/B students being given a chance on the higher paper as a 'second bite at the cherry'.

Table 6.2 Student reasons for prefering either of the two models of 2-tier examinations

Pilot model		Trial model	
n = 121	28% of sample	n = 298	69% of sample
The easier questions (on Trial) threw me		More enjoyable	
Too many easy questions (on Trial papers)		More relaxed, more time to revise	
Because it is what I studied		Fairer	
More time to prepare/do the papers		It was easier, the other was too hard	
We had done more revision/practice		I had a better chance of getting a higher grade, a pass, bigger range of grades	
		Good/ wide range/ mix of questions	
		I felt challenged/ more confident/ suited my ability better	
		Easier questions at the start built my confidence	
		I liked the style of the questions and the exam better	

22 questionnaires (3%) gave no response or were uncodable

In relation to this examination the experience of those students predicted a B or C grade are also interesting. All students take the central 'C/D' grade paper then there is a choice between the A*-B paper and the E-G paper (see Fig. 1). Evaluations of the Pilot to date reveal teacher concerns about the negative experience of candidates predicted a B for whom typically two-thirds of the A*-B paper are inaccessible. Similarly, decisions need to be taken about students predicted a C grade: the G-F paper is likely to be experienced as trivial while the A*-B paper will be very difficult. Table 6.3 (below) indicates that, whatever their predicted grade students preferred the Trial papers.

Table 6.3 Student preference for Trial or Pilot by predicted GCSE grade on Pilot

	Predicted GCSE grade							
	A		B		C		D	
	Pilot	Trial	Pilot	Trial	Pilot	Trial	Pilot	Trial
n	27	61	31	42	41	108	12	46
%	30	68	41	55	27	73	21	79

- Grade totals that do not add to 100% are the result of students expressing no preference either by omission or by saying so.
- Only 29 students were predicted a grade E, F or G – their comments have been excluded as the numbers are so small

It is notable that the proportion of students predicted a B grade and preferring the Trial is lower than other groups of students. Generally the reason given for the stated preference related to the degree of difficulty experienced with the papers – these comments were evenly split between the two models. Of those preferring the Trial the most frequent comment was that these papers offered a better range of questions. Fewer students provided reasons for preferring the Pilot, and the most frequent reason given was that they preferred the style of

the questions or the paper and that it felt fairer, it is also possible that the greater familiarity with this examination structure helped them to feel more comfortable with it.

Of those predicted a grade C, 68% were entered for the Foundation tier in the Trial and 32% were entered for the Higher papers.

Table 6.4 Preferences of students predicted a grade C

	Entered 2-tier Trial Higher		Entered 2-tier Trial Foundation	
	Pilot	Trial	Pilot	Trial
n	20	27	21	81
%	43	57	21	79

Preferences amongst those who sat the Higher Trial papers are fairly evenly balanced (given the small sample size). The difference is more marked for those sitting the Trial Foundation papers where there is a strong preference of the Trial papers. The reasons given were the same as those given by the students predicted a B.

The C grade is available at the extreme top end of the range of potential grades for Foundation students on either of the two tier models (although it is not the lowest grade available on the Higher tier, see Fig.1). We speculate that attaining a grade C may feel fairly precarious and difficult. In the light of this, the Pilot model provides one ‘easy’ and one ‘hard’ paper whereas the Trial provides two more balanced papers – this may well feel ‘safer’ to students.

7. Other student comments

Like the teachers, the most frequent comments related to the time taken to do the Trial papers (or the length of the papers) – it was felt that either the papers were too short, or the time was too long.

Across the two models there were no significant differences between the responses of girls and boys beyond what we might expect from previous research – for example, girls are generally happier than boys doing coursework, boys are generally more confident than girls about the amount and kind of revision they have done. We do not report these findings in any detail here.

A surprising finding was the degree to which students identified with the three-tier 'sets' they had been taught in. While this was true for all the groups it was particularly marked and noticeable amongst the 'Intermediate' students many of whom expressed concern about the loss of this middle ground. In interview they were clear which way they would have moved if the change had come before they left school with some opting for the perceived safety of the Foundation tier and others saying they would have enjoyed being more challenged and would have been keen to move to the Higher tiers and have a go at the higher grades. However, within the Intermediate set they felt noticed and special – it was important to them that they were neither 'too clever' nor 'thick' (the iconic /stereotypical positions available for learners in maths classes) and that this had been recognised and publicly sanctioned with a special examination.

7.1 Students taking the Trial and 3-tier traditional examinations

Responses from 1539 student questionnaires were coded onto Excel. The coding used is included in Appendix 7.

Overall the students preferred the 2-tier examination (59.8% to 40.2%). This holds true whatever grade students were predicted (Table 7.1)

Table 7.1 Preferred examination model by predicted GCSE grade

Predicted GCSE grade													
A		**B**		**C**		**D**		**E**		**F**		**TOTAL**	
3-tier	2-tier	3-tier	2-tier	3-tier	2-tier	3-tier	2-tier	3-tier	2-tier	3-tier	2-tier	3-tier	2-tier
34%	66%	45%	55%	46%	54%	27%	73%	45%	55%	46%	55%	40.2%	59.8%

However, if we look at the preferences in relation to the examination tier they had been preparing for we get a slightly different picture (Table 7.2 below)

Table 7.2 Exam preference by tier prepared for

	3-tier examination entered							
	Higher		Intermediate				Foundation	
	2-tier examination entered							
	Higher		Higher		Foundation		Foundation	
	examination preferred							
	3-tier	2-tier	3-tier	2-tier	3-tier	2-tier	3-tier	2-tier
n	154	367	246	114	103	248	38	81
%	30	70	68	32	29	71	32	68

In all cases the two-tier Trial examination was clearly preferred with the exception of the Intermediate students who sat the Higher papers for the Trial where the proportion preferring the 3-tier was highest. This may seem surprising until we look at the reasons they gave which show that they felt ill-prepared for the Higher paper the A and A* material being unfamiliar. This was the only group for whom the paper may have contained surprises. The students we interviewed were all sanguine about this realising that it was an unfortunate result of the way the Trial had been set up and not a reflection on them:

> "There was no Intermediate for the two-tier which meant the papers were too easy or too hard"
>
> "The two-tier exam contained some questions we had not covered in class like vectors"
>
> "I think I could have done a lot better on the 2 tier exam if I had been taught the higher tire syllabus however, because I hadn't learned many of the things on the paper it was impossible for me to do some of the Qs"
>
> "Simply because I felt really confident when I turned every page of the [3-tier] exam! I did also like the 2 tier exam but I ticked the 3-tier because I'm very confident that I've achieved my target."
>
> "Better feel, more choice. Gives everyone a chance to achieve well. Better all round."(Trial)
>
> "I thought some questions [on the 2-tier papers] were very easy, which confused me and so I looked for a more complicated method therefore I may have got easy questions wrong as I expected the paper to be too hard for me."
>
> "I disliked all of the questions [on the 2-tier papers] and found the exam very stressful and it made me feel really stupid."

At interview many of those Intermediate students who sat the Higher Trial papers talked about the difficulty and sense of injustice of seeing questions at the end of the paper that carried so many marks but that they could not attempt.

Students who preferred the Trial liked the potential access to higher grades (including a C for Foundation students), many found the paper easier than they had expected and liked the lead in provided by the easier questions.

7.3 Views on teaching and learning since KS3 SATs

We asked all students to respond to two questions relating to their learning and progress:

- Most of the time did your maths lessons feel: too easy, about right, too hard?
- Do you feel the progress you have made since KS3 SATs has been: very good, good, very little, no progress?

We report the results broken down by predicted GCSE grade (none report being predicted G or U).

It is important to remember that the Trial does not show up in these tables and that those taking the Pilot syllabus have no experience of the 3-tier examination structure.

Table 7.3: Student perceptions of teaching and learning

Predicted grade		A		B		C		D		E		F	
		Pilot	3-tier	Pilot	3-tier	Pilot	3-tier	Pilot	3-tier	Pilot	3-tier	Pilot	3-tier
	n =	90	329	76	383	150	520	58	214	25	48	2	12
Mostly lessons felt …	too easy	1	8	0	8	11	4	9	9	4	6	0	9
	about right	94	88	92	81	83	85	81	73	92	71	100	73
	too hard	4	5	8	11	7	12	10	18	4	23	0	18
My progress since KS3 SATs has been …	very good	37	35	32	23	20	15	23	11	24	6	50	0
	good	63	59	61	65	72	71	56	60	48	60	50	58
	very little	0	5	8	12	7	13	18	26	28	32	0	42
	no progress	0	0.6	0	0.8	0.7	2	4	3	0	2	0	0

While we need to be cautious in our interpretation of students' reporting of levels of satisfaction with teaching and their progress, there would appear to be some drop in satisfaction below B grade for teaching in Pilot schools, and below C for 3-tier schools. This might relate to the importance of the two grades within the two structures – for those taking the higher Pilot route the C grade is talked of as a given – something they 'bank', whereas for those taking the 3-tier examination the C is the key grade and schools focus on getting as many students as possible to the grade C boundary.

In terms of their perception of their progress since KS3, those who followed the Pilot programme are generally more satisfied but the effect is small for those expected to attain higher grades. Students' sense of progress drops off below a C for 3-tier

7.4 Plans to stay at school and study A level

We asked students whether they were planning to continue studying post 16 and whether maths had ever been part of those plans. Unfortunately we did not have time to fully analyse the qualitative data relating to these questions.

Of those who reported having thought about carrying on with maths at A level, 17% of the Pilot students and 11% of those sitting the traditional 3-tier examination were taking the Foundation tier papers. They may have been reporting having thought about carrying on with mathematics many years ago, or they may have unrealistic expectations. If this greater proportion is due to unrealistic expectations it would confirm teacher concerns raised in relation to the structure of the 2-tier examinations.

The breakdown by gender shows the numbers of boys and girls considering maths at A level to be roughly equal with marginally more boys than girls having thought about it.

Table 7.4 Student plans post 16

Students who had been studying:		QCA/OCR Pilot		3-tier maths		Total	
		n	%	n	%	n	%
Are you planning to stay on at school or college?	Yes	389	90.5	1418	90.9	1807	90.8
	No	41	9.5	142	9.1	183	9.2
Have you ever considered studying mathematics at AS or A2 level?	Yes	143	33.3	460	29.6	603	30%
	No	286	66.7	1096	70.4	1382	70%

However, if we look at the breakdown by gender a marked difference occurs with significantly more girls from the Pilot route considering maths at A level than girls from the 3-tier route (Table 7.5). Since this compares the Pilot with the 3-tier examination and there is no information about the Trial model we need to be cautious about our interpretation. However, it is well established that girls are over represented at Intermediate level and are therefore restricted in what they can achieve (it would not generally be possible to move onto an A level course from an Intermediate syllabus). Entry patterns in the 2-tier Pilot examinations suggest that, forced to commit girls predicted a B to the Higher or Foundation tiers, teachers will give them access to the higher papers and that, given this access, more girls will continue to study mathematics. It is not clear whether this would also be true in the 2-tier Trial model but, at this stage we can only assume it would be.

Table 7.5 Plans to study A Level by exam structure taught and gender

		QCA/OCR pilot		Traditional 3-tier	
		n	%	n	%
Those who have considered studying mathematics at AS or A2 level	Girls	74/217	34.1	156/670	23.5
	Boys	69/213	32.5	304/905	34.0

8. Summary of main findings

This section pulls together the findings reported separately in sections 2-7 and looks at the implications for the Pilot and Trial models.

We have organized our analyses around two central questions:

1. ***What are the main purposes of GCSE mathematics examinations*?** The tension here is between certification of 16 year olds finishing compulsory schooling (accessibility/ inclusion) and preparing and identifying higher achievers who may study GCE mathematics (stretch).
2. ***What does a GCSE mathematics grade mean*?** This examines the inferences about mathematical competencies that can be drawn from a particular grade. Grade B is of particular interest as it often the minimum entry grade for GCE mathematics. It also involves *comparability* issues in moving to a two-tier model. In relation to *validity* we looked at whether papers have functioned as intended.

In addition there are a series of implementation issues that we reported on from the analyses of teacher and students surveys and interviews. These include how teaching will be organized, progression issues and resources.

In reading and interpreting these findings it is important to be aware that the short notice at which the Trial was run may have implications for the robustness of the data although we generally deal with it as if it were unproblematic. For both the schools and the exam boards, this is the only time they have run this form of two-tier mathematics GCSE examination – there was no opportunity to trial the Trial. Perhaps more importantly, it was only in March/April 2005 that schools signed-up for the Trial; this left very little time to prepare students for additional material they might encounter or the new structure of the papers. Despite this students and teachers were extremely enthusiastic about the opportunities the Trial offered.

8.1 What are the main purposes of GCSE mathematics?

The current structure of GCSE mathematics, with much of the content linked to specific grades, makes it difficult to fully achieve both accessibility and stretch. Though both Trial and Pilot structures offer all candidates the chance to gain a grade C, we see the two models as offering different emphases. The Pilot seeks to offer stretch through a higher paper which targets only grades A*-B. While these candidates also take a C-D paper, this is discounted for those with higher grades. The Trial offers a broader approach in assessing A*- D level work on the same paper, all of which contributes to the final grade. This limits the proportion of higher grade material and reduces proportion of A* - A questions (see next section).

Accessibility and inclusion vs. demand.

The impact of these different paper structures on the students taking them is important as it may affect their exam performance and colour future attitudes to mathematics. This is particularly important as this will be the last formal mathematics examination that most students take. Our survey findings showed that **overall 60% of Trial students preferred it to the three-tier examination** – even though they had had little or no time to prepare for it. However when this was broken down by expected grade and by tier of entry (Table 7.2) the picture is more complex, with those predicted a grade B on the three-tier Intermediate papers preferring the three-tier examination structure (on which they would have not had to do A*-A demand questions, for which they had not been prepared).

We had survey returns from 430 students (in five schools) who sat both the Pilot and Trial examinations. **Two thirds of these students preferred the Trial to the Pilot model.** The reasons they gave for this were generally related to the range and accessibility of the questions and the confidence-building effects of easier lead-in questions (Table 6.2). While two-thirds of those predicted a grade A preferred the Trial, only 55 per cent of grade B students did, as did only 57 per cent of grade C candidates who were entered for the higher tier (compared with 79% of grade C candidates entered for the foundation tier). We speculate that this may be because these candidates felt the C-D paper offered them security, a paper they did well on, so the higher paper allowed a low risk chance of a higher grade.

The mathematics teachers in the eight schools that took both the Trial and Pilot examinations were divided in their preferences. For those who preferred the Pilot this was often because it offered more challenge for higher achieving students and allowed for more targeted teaching of specific topics. Those favouring the Trial felt it was a more positive experience for the students – they did not have to sit 'inaccessible' papers and all their work contributed to the final grade.

The OCR examiners who have been involved with the Pilot for the last three years **preferred the Pilot model** to the Trial, for which they were also responsible in 2005. This was essentially **because it provided a more demanding paper for the higher achieving candidates**.

The most consistent negative comment about the Trial papers was the time allowed (2 hours), which both teachers and students thought was too long, as most candidates finished well inside this time. While this can be easily remedied in future either with a shorter time or more questions, examiners at the OCR awarding meeting commented on the constraints on paper setting (e.g. the number of questions they could set) of the 'Targets and Tariffs' document used to monitor paper setting.

Even though the Trial examination was experienced as more accessible, the awarders graded the Trial papers slightly more severely. The Trial grade distributions were generally slightly lower across each of the awarding bodies and the OCR Pilot than the students' three-tier results across all the awarding bodies. To compensate for this our analyses brought the overall Trial grade distribution into line with both Pilot and 3-tier examinations.

The comparison of the adjusted **Trial and three-tier grade distributions showed that only around 70% of candidates received the same grade on both examinations** (Appendix 8). The implication of this is that, while a move to 2-tier examinations may produce the same overall grade distributions, at an individual level, there may be grade differences for a minority of students. This is perhaps an inevitable outcome of any change in examination structure.

When the adjusted Trial and Pilot results were compared to the 'fine grade' KS3 results, we found that **for those candidates who had achieved levels 2-6 at key stage 3, the Trial offered a better chance of a higher grade. It was only for students gaining level 7 or above that the Pilot was likely to confer some advantage although most of these candidates did equally well on both models** (Figure 3e).

While we need to be cautious, **there is evidence that a move to a 2-tier examination structure will encourage girls to continue studying mathematics at A level** (Table 7.5). It would appear that past evidence that girls are over-represented in the Intermediate tier may have been depressing their access to higher level courses and that a move to 2-tier examinations will remove this 'glass ceiling'.

8.2 What does a GCSE mathematics grade mean?

We have used this question to address issues about what information a GCSE mathematics grade carries, what can be inferred about a student's mathematical competence, and what may be the threats to the validity of grading.

Both models use a compensation approach to grading in which the grade on a paper is determined by the total mark rather by meeting particular criteria (e.g. 'must be able to…'). In order to provide more information about competencies, the Pilot has restricted each paper to narrow range of grades. Unusually for GCSE, curriculum content in GCSE mathematics is directly related to specific grades. The assumption underpinning the Pilot model is that questions can be accurately targeted at precise levels of difficulty, even though the examinations are not pre-tested (unlike national curriculum tests). This is in part because only, for example, grade A*-B students will have taught certain questions.
We investigated whether the level of demand of questions functioned as intended. **We found on the Pilot that our question level data (Section 3) challenged the assumption that particular questions could be precisely targeted at particular grades**. We would have expected questions on the A*-B paper to have been consistently more difficult than those on the C-D paper which would in turn have been more difficult than E-G questions. This was often not the case. After adjusting for which students had attempted the questions, we found that four of the questions on Paper 2 (targeted at grades C&D) we harder than 10/25 of the questions on paper 3 (targeted at grades A*-B) (Figure 3f). Similarly, there were six questions on the C-D paper that were easier than at least two questions of the questions on the E-G paper.

The question level data from the AQA Trial showed similar variability. Targeting is less precise in the Trial because questions are designated 'low', 'medium' or 'high' tariff as opposed to being 'grade specific' in the Pilot. There is also a broader range of questions on each of the Trial papers. For both of these reasons the Trial is less dependent on the precise targeting of questions so the threat to validity is not as great.

What information does a grade carry?

What can be inferred from a grade B has been an issue for the three tier model because a grade B can be achieved from both the higher tier (A*-C) or the intermediate tier (B-E), with the tiers differing in the content studied. Both the two tier models resolve this by having only one route to a B, although there remain two routes to a grade C.

In the case of the Pilot a grade B can only be achieved through taking the A*-B paper, which 'guarantees' higher level content (the basis for higher work, for example at A level). This is not necessarily the case with the Trial since 55 per cent of the weighting was at C-D level of demand (the AQA grade B boundaries were set at 50 and 40 per cent; OCR's at 49 and 55).

However even though, in 2005, 50 per cent of the Pilot paper was targeted at grade B, the grade boundary was set at 24 per cent, having been 18 per cent in 2004 and 14 per cent in 2003. There are reliability issues here, as well as public credibility ones. A grade based on less than a quarter of the marks allows only limited inferences about what a candidate knows and can do.

This problem is compounded by the breadth of the grade B range of marks from the Pilot papers. While Pilot candidates who got 24 per cent of the higher paper marks got a grade B, so did those who gained 55 per cent – a very different level of performance which could not be inferred from the grade. For Trial candidates the 'B band' is narrower covering around 20

marks in the OCR Trial (Table 4.3). **While the Pilot claims to carry more information about the skills and knowledge associated with a particular grade, it would appear that this is questionable.**

Paradoxically, it was the students with a predicted grade B who had the largest minority preferring the Pilot (41%) to any grade (Table 6.3).

For the Trial model there remain two routes to grade C, which raises well rehearsed comparability issues. While the inclusion of common questions across the two tiers has sought to address this, there are still issues about the comparability of performance. AQA set the C boundaries of the higher tier at 34 and 25 per cent, which allows limited inferences about competencies. It is however an improvement on the three tier model, for which the boundaries were 21 and 22 per cent. The Pilot model avoids this problem by a single route, though a grade boundary of 39 per cent, on a paper on C was the top grade, was well short of the intended threshold mark.

8.3 Implementation issues

Teachers did see not moving to a two tier model as particularly problematic in terms of classroom organization, setting and resources:

- The Trial model was likely to raise coverage issues as students would have to be prepared for a wider range of topics.
- Some teachers welcomed the Pilot because it allowed them to focus strategically on a narrower range of higher level topics.
- Some teachers raised tactical issues of what topics they should teach students who may just be getting a grade B.

The teacher perception that the Trial was more accessible to students led to the concern that this could encourage students, for example those with grade B, to take GCE A level mathematics courses on which they might struggle.

Teachers were most negative about GCSE coursework which they believed to be time-consuming and increasingly unreliable because of the ease of accessing model answers.

8.4 Conclusion

The move to a two-tier GCSE mathematics examination has been widely welcomed by teachers and students. The two models we have evaluated have much in common in terms of curriculum and the accessibility of the iconic C grade for all candidates. If one model is to be chosen, we believe the key questions are about the main purposes that the GCSE mathematics serves and what is expected that can be inferred from a grade. We see a different emphasis in relation to inclusion and stretch between the Trial and the Pilot. We have questioned, for each model, what information a grade carries. Our evidence suggests that grade based inferences about mathematical competencies have to be extremely cautious, particularly as targeting the difficulty level of questions and papers proved difficult for both models.

The strength of the Trial is its accessibility, that of the Pilot is the apparent stretch provided for those who gained level 7 at key stage 3 SATs. The obverse of this provides the risks: the

Trial provides lower demand for the highest achievers, and the structure of the Pilot makes two of the papers potentially inaccessible for lower achievers (CD paper too hard for E,F,G candidates, A*-B paper too difficult for C/D candidates) while the C/D paper is too easy for the A*-B candidates.

---

9. References

Smith, A. (2004). *Making mathematics count: The report of professor Adrian Smith's inquiry into post-14 mathematics education*. London.

Working Group on 14-19 Reform. (2004). *14-19 curriculum and qualifications reform: Interim report of the working group on 14-19 reform (Tomlinson group)*.