Standards in A level Mathematics 1986-1996

John F Bell, Tom Bramley and Nicholas Raikes
University of Cambridge Local Examinations Syndicate, 1 Hills Road
Cambridge, CB1 2EU

Paper presented at the British Educational Research Association Annual Conference
(September 11-14 1997: University of York)

Abstract

This paper will present the results of an investigation into the ways in which the UCLES A level mathematics syllabuses have changed over time. It will include a discussion of the changes that have influenced this subject over the period in question and consider the difficulties of interpreting the changing distribution of A level grades over this period in time.

Details of a study involving archive scripts from 1986 and 1995 will be given. This study involved the scrutiny of scripts by experienced examiners to find if there were any changes in how grades were awarded.

Introduction

Recently there has been a debate about the increasing proportion of candidates obtaining grade A's for mathematics A-level. It has been argued that this improvement is not the result of a genuine improvement in the quality of candidates but that it is an example of grade inflation which has been defined as an increase in grades without a concomitant increase in ability (Ziomek and Svec, 1995). The Institute of Mathematics and its Applications, the London Mathematical Society and the Royal Statistical Society (LMS/IMS/RSS, 1995) produced a report *Tackling the Mathematics Problem* which suggested the following:

4A Students enrolling on courses making heavy mathematical demands are hampered by a serious lack of essential technical facility - in particular, a lack of fluency and reliability in numerical and algebraic manipulation and simplification,

4B Compared with students in the early 1980s, there is a marked decline in students' analytical powers when faced with a simple two-step or multi-step problem;

4C Most students entering higher education no longer understand that mathematics is a precise discipline in which exact reliable calculation, logical exposition and proof play essential roles; yet it is these features which make mathematics important.

Sutherland and Pozzi (1995), in a study of the nature and difficulties which undergraduate engineers are experiencing with mathematics, obtained questionnaire data from 42 engineering and mathematics lecturers from four universities. They concluded that the decreased emphasis on algebra and the almost complete removal of Euclidean geometry from the pre-16 curriculum has resulted in very little emphasis being placed on mathematical deduction. At the same time, some researchers have suggested that mathematics and sciences were difficult compared with other A-levels (Fitz-Gibbon and Vincent, 1994).

Because of the concern expressed about the standards in A-level, OFSTED/SCAA (1996) were instructed to prepare a report to find out what had been happening. They compared scripts from 1985 with scripts from 1996. Mathematics was one of the subjects considered in this report which came to the following detailed conclusions:

- in pure mathematics, standards set at the grade E boundary have fallen in two of the three syllabuses scrutinised;

- at the grade A boundary in pure mathematics, whilst many aspects of performance were comparable with 1985, candidates in 1995 were not required to demonstrate as much competence in the important areas of problem-solving, reasoning and algebraic manipulation;
- standards have risen in the statistics syllabuses reviewed;
- evidence about standards in mechanics is inconclusive.

This report somehow managed to summarise the findings as follows: "It is probable that, over 20 years, standards shown by the majority of candidates have risen, because of the greater emphasis on breadth and the accessibility of questions, but they have fallen for the most able, owing to changes in syllabus content, a different emphasis on depth of treatment and greater structuring of questions."

This paper will report the use of Thurstone paired comparison methodology to investigate what has happened with A-level mathematics standards. This study used small samples of scripts from 1986 which had been loaned for research purposes to Southampton University. This paper presents some preliminary results from a report into this issue which is in preparation and will only address some of the issues relating to standards in A-level Mathematics.

Changes over time

Some of the perceived problems arising in assessing changes over time come from a failure to distinguish the differences between performance and expectations. For example, in an 1863 examination for junior candidates, in a section on preliminary arithmetic, the candidates were asked to answer the following question:

By practice, find the rent of 63 ac. 3 ro. 27 po. at 30s. an acre.

It is reasonable to assume that modern A-level mathematics candidates would not be able to work out that the answer was £95 17s. 6¾d. because of their unfamiliarity with the units. There almost certainly has been a decline in ability to perform calculations involving pre-decimal coinage and imperial measures but this is not surprising and is of no practical importance.

The technical issues in detecting a change in performance on a skill are relatively straightforward but the interpretation of these changes is not. This is much more subjective and some people would be more concerned about a decline in the ability to do algebraic manipulations than others. The content of mathematics A-level syllabuses require a consensus of opinion. Reid (1991) argued that 'if policies on access at 16+ are to be chosen, rather than be dictated by the inertial influence of ideology, then Britain must follow other countries in treating this stage of education as an integral part of the system as a whole and apply rules for transfer to university which are in the interests of all parties and not just teachers in higher education.'

Because the consensus of opinion about the relative importance of component skills and knowledge within a subject area change over time, it is not unreasonable for these changes to be reflected in the content of A-level syllabuses and the construction of examination papers. This means that changes in some areas are to be expected.

To discuss changes over time, it is necessary to define carefully what is meant by a change in standards. In the table below, different types of changes have been identified.

Type of Change	Description
Change in coverage	This refers to changes in the content of the syllabus and the range of material covered. This will also be taken to include changes in the relative importance of material within the syllabus.
Change in accessibility	This refers to the relative difficulty of obtaining marks in an examination. An examination with a mean percentage mark of 50% would be much more accessible than a examination with a mean mark of 25% given the same candidates.
Change in grading standard	This occurs when a candidate with a particular level of mathematical achievement would obtain different grades on different occasions.

Changes in coverage and accessibility do not necessarily mean that there has been a change in grading standard. Given that even the best candidates are not expected to obtain full marks in A-level examinations, a syllabus may contain material that only a small minority of the most able grade A candidates master. If this material were omitted, there would be a change in coverage and a change in accessibility since without this material the mean percentage mark would rise. However, by adjusting the pass mark there might be no change in grading standard.

Changes in coverage are a major source of difficulty in interpreting changes over time. If the emphasis on a particular part of a syllabus is reduced, then it is likely that candidates will spend less time studying it with a commensurate decline in their understanding of the material. If this material is vital for a course of higher education, then the lecturers of such candidates will conclude that there has been a serious decline in standards. The fact that other parts of the syllabus have been given greater emphasis with a resulting increase in understanding may not be noticed or appreciated.

The above changes are not the only possible explanations of changes between distribution of grades for a particular syllabus. There can also be changes in the quality of candidates who choose to take the A-level examination. This can operate at both the syllabus and the subject level. At the syllabus level, an improved grade distribution could result from centres which tend to have weaker candidates entering them for another syllabus in the same subject which is considered more appropriate for their candidates. At the subject level, the range of subjects offered by centres and examination boards has increased which means that a subject could lose the candidates who are less interested and/or less able to other subjects. This problem would arise even if there was a single examination board offering only one syllabus in each subject.

Background to the debate on declining standards

Before considering the changes in A-level mathematics, the changes in overall A-level entry, mathematics education and higher education will be considered. In particular, it is necessary to investigate whether there is any evidence to support an improvement in quality of the candidates which would be necessary to explain the changes in the grade distribution if there had been no change in grading standard.

Over the period of the study there has been as massive decline in the number of seventeen-year-olds. The cohort size dropped from 748,200 in 1985/86 to 556,700 in 1993/94, a decline of 29% (DES, 1994a). Naive commentators have failed to appreciate the magnitude of this change and this has resulted in much sensationalist reporting. This decline has been offset by an increase in the percentage of seventeen-year-olds studying A-levels. This percentage has increased from 20% in 1985/86 to 34.5% in 1993/94. This has resulted in a net increase in the numbers of seventeen year-olds taking A-levels.

In the period under consideration, there has been a large decrease in the cohort (of course, it should be recognised that some A-level candidates do not come from the cohort of seventeen-year-olds). If it is assumed that in 1985/6 the entry for A-level courses included most of the seventeen-year-olds with the potential to obtain good A-level grades then these changes in uptake would point to a dilution in the quality of sixth form entries. The situation relating to mathematics is more complex. Despite the increased percentage uptake of science and mathematics A-levels, the number of candidates taking combinations of *only* science and mathematics A-levels had declined by 20% from approximately

93,500 in 1985/6 to 75,500 in 1993/94. However, this represents a slight increase from 12.5% to 13.3% of the cohort DES(1994b). This could mean that mathematics candidates are less likely to have been studying science A-levels. If studying science reinforces mathematics skills, e.g. physics and the mechanics sections of mathematics papers, then this could be taken as a reason for expecting poorer performance.

Goldsmith (1995), deputy chairman of the School Mathematics Project, and chief examiner of the Oxford and Cambridge Schools Examination Board's non-modular A-level, argued that the clientele for A level examinations changed and this meant that the examinations had to change too. He suggested that the "Mathematics assessment level has become more accessible to weaker students (that is, easier); the motivation of students has increased; passmarks have risen; and failure rates have been further reduced."

In the same period, there has been a dramatic increase in the number of students in higher education. Between 1985/6 and 1994/5 the number of students in higher education increased from fewer than one million to more than one and half million students. However, more than 40% of this increase has been by students with qualifications other than A-levels. The transbinary data published by HESA (1995) show that in 1987 there were 41% good degrees (i.e. 1st's or 2.1's), in 1990 46%, in 1991 47%, and in 1992 48%. In the old universities, the number of Firsts awarded increased systematically from 6% of students in 1981 to 10% of students in 1993. The HEQC (1995) found that in every subject studied there has been an increase in the number of good degrees because 2.1's have now become the modal degree class across the system. Only considering percentages ignores the effect of the growth in student numbers. This makes the changes more dramatic, for example, in 1995 more students received 1st's and 2.1's than received 1st's, 2.1's, 2.2's and 3's in 1986. These simple statistics, of course, mask the differences between subjects, institutions and types of institution, and the increase in numbers of female students. Since higher education institutions would not risk their reputations by allowing their standards to fall this would provide evidence in support of an improvement at A-level. However, Kahn and Hoyles (1997), in a case study of single honours mathematics in England and Wales, found that the range of mathematics had broadened away from traditional pure mathematics, the advanced content covered in the three year degree had been reduced, and assessment had been changed with more structured questions and more calculation at the expense of proof. They concluded that these changes led to an overall reduction in the rigour and depth of degrees.

There is evidence of changes in attainment in mathematics in education prior to A-levels. The Assessment of Performance Unit had as one of its objectives the measurement of changes over time. Unfortunately this Unit was scrapped when the National Curriculum was introduced. There is, however, some relevant evidence. Foxman et al. (1991) found that the performance of eleven-year-olds on questions about fractions, computations, applications, rate and ratio declined significantly between 1982 and 1987. The pupils who took part in these surveys would have reached sixth-form age in the period of the study.

The period 1986 to 1996 has been one of great change in education. In this period, there have been many changes in legislation which were intended to raise educational standards. These included the change from GCE O-level and CSE examinations at 16 to the unified GCSE examination, the creation of the National Curriculum and its associated assessment and the introduction of common cores for A-level subjects.

The change from GCE/CSE to GCSE means that changes in performance at age 16 are difficult to interpret. GCSE examinations are not necessarily intended to assess the same skills and knowledge as their predecessors and cater for a wider ability range which means that changes in grade distribution have no simple interpretation. One important change in GCSE has been in the structure of science education. Many candidates for mathematics take double science GCSE instead of Physics as a separate science O-level. This means that sixth formers will have studied less Physics prior to starting their A-level studies which could influence their performance on mechanics question at A-levels.

Sutherland and Pozzi (1995) compared the syllabuses of 1993 mathematics GCSE courses with those of 1983 O-level courses. They found that there had been an overall reduction in content in the move from Scottish Ordinary grade to Standard grade and O-level to GCSE and that this was particularly marked in the areas of trigonometry and algebra. They noted that the emphasis in pre-16 mathematics is on the solution of mathematics situated within practical problems. This means that if algebra is used in this type of problem, the algebraic formulation can often be solved by trial and error.

There has been some suggestion that GCSE mathematics is not as good a preparation for A-level as O-level. For example, Wiliam et al. (1995) considered the size and nature of the gap between GCSE and A-level Mathematics. A group of eight teachers and researchers looked at examination papers, marking schemes and candidates' scripts for both GCSE and A-level, for a range of examination boards. In a very limited sample of scripts they found that algebra was disproportionately difficult for most candidates and there were examples of candidates obtaining a grade A at GCSE without answering correctly any algebra questions. When they asked experienced sixth form tutors for the main reason for lack of success, lack of algebra skills was given as the most popular reason.

The evidence from this section is mixed. The change in proportion of candidates entered for A-level mathematics suggests that quality of the entry should have declined. There is evidence to suggest that candidates are now not as well prepared for some aspects of A-level courses. However, the results from higher education would support an improvement in performance and anecdotal evidence would support a decline.

Comparison of Syllabuses

Three UCLES syllabuses were considered in the this study: Mathematics Syllabus A 9200 in 1986, Mathematics Syllabus 9205 in 1986 and 1995. This study was made possible because of the availability of two small samples of scripts from 1986 which had been loaned to Southampton University for research purposes. 1986 was the first year that common cores at A-level were used in Mathematics. The content of the 9205 syllabus was the same in 1986 as in 1995 except that in 1986 there was an extra topic in section B of Paper 1 - matrices. Although the 1996 data were available at the time of the study, this was not used for comparing scripts. This was because in 1996 the 9205 syllabus was replaced with one which no longer offered a choice of questions. The structure of the 9205 1995 examination was not changed from that of the 1986 9205 examination. The change in structure over 1995/96 was monitored both internally by examination board staff and by the office of Her Majesty's Chief Inspector of Schools (OFSTED, 1996). No evidence of a change in grading standard was found by either.

Although there had been no change in the 9205 syllabus, there had been changes in the style and the layout of the paper 2. This is illustrated by enumerating certain features of the questions as they appear on the paper. In Table 1 the mechanics questions in 1986 and 1995 are compared in terms of how the questions were broken down into sections with number of marks allocated. In addition, the number of questions where an answer was explicitly asked for was counted. The length of the question was measured by counting the number of lines of text for the question. For example, question 1 from the 1986 paper had two parts that were each worth seven marks. Two explicit questions were asked in each part. The first part of the question was six lines of text long and the second part of the question was five lines of text long. The question also included a diagram.

Table 1: Question structure for Mechanics questions Mathematics 9205 1986 and 1995

	1986			1995		
	marks for each part (visible on paper)	number of questions asked	lines of text in question	marks for each part (visible on paper)	number of questions asked	lines of text in question
Q1	7,7	2,2	6,5 diag	3,7,2,2	1,2,2,1	4,5,2,1 diag
Q2	8,6	2,1	5,1	3,5,3,3	2,1,1,1	3,5,1,2
Q3	8,6	1,1	6,4	6,3,3,2	1,2,1,1	9,2,1,1 diag
Q4	4,5,5	1,2,1	4,2,1	5,3,6	1,1,1	3,3,4
Q5	4,10	1,3	4,4	6,3,5	1,1,1	5,1,5 diag

*'Number of questions asked' is a count of each answer the candidate is explicitly asked for. For example, "Find the magnitude and direction of the force" would count as 2 questions.

From the table it is clear that the questions were more structured in 1995 than in 1986. For example, there are three diagrams in 1995 and only one 1 in 1986. The number of visible marks had risen from 11 to 18 and the number of explicit questions had risen from 17 to 21. More questions were broken down in terms of the marks given on the paper in 1995 and a few more questions were explicitly asked in 1995. The amount of text increased from 42 lines to 57 lines.

This analysis was replicated for the statistics and pure mathematics sections of the paper and the same pattern of changes was observed.

Pure Mathematics

The pure questions changed the least over the 9 years and appeared to be similar in structure and content. This is not unexpected given the introduction of the common core and that the syllabus was virtually unchanged.

Mechanics

The style of the mechanics questions in 9205 changed in 1988. In 1986 all the questions were in terms of symbols. A calculator would not have been needed. In 1995 most questions contained numerical values for all terms except those the question was asking to be calculated. This was done because an inter-board comparability study found that the UCLES questions were harder than those offered by the other examination boards. Although this was an attempt to make the examination more accessible there was no intention of having a change in grading standards.

Some research is being carried out into the performance of current candidates on 1986 questions but it was not possible to complete this study concurrently with the examiner scrutiny study. In this planned research, the change from formulae to numbers in mechanics questions will be investigated. For the 1986 questions, the formulae were replaced with numbers and for the 1995 questions, the process was reversed. Obviously, this cannot be done for all questions. Although this process proved to be relatively easy, there were problems in amending the mark schemes.

Statistics

For the statistics questions, the topics covered are identical and the style of the questions is very similar. The 1995 paper seems more varied and interesting, with more attempt to give the questions a relevant context (e.g. vegetable crops) rather than a very artificial context (boxes of red & green apples; random arrivals of broken noses at a surgery). In contrast, the 1986 paper relies heavily on probability questions. Unlike the mechanics questions, both papers contain questions that would require a calculator.

Grade Distributions

The grade distributions for UCLES linear mathematics syllabuses are given in Table 2. In 1995 there were also entries for the new UCLES modular syllabus which have not been included in the table. The entries for the UCLES linear mathematics syllabuses had declined by 2869 candidates (36%). As noted earlier, the cohort has also declined in size over the period. Given the cohort size had declined, a decline in entry would be expected but for the fact that the overall entry for mathematics for all boards had increased. There has been a large increase in the cumulative percentages of candidates, e.g., in 1986 72.6% of candidates obtained at least a grade E and in 1995 85.7% of candidates obtained at least a grade E. It should be recognised that the C and D boundaries are not determined by examiner judgement but are calculated arithmetically. This means that if the distribution of marks changes there can be changes in the percentages obtaining grades C and D which are beyond the control of the awarding committee.

Table 2: Grade distribution for UCLES linear mathematics syllabuses

Grade	1986 UCLES Syllabuses number	Cum %	1995 UCLES 9205 number	Cum %	Difference
A	1458	19	1544	31	12
B	1315	35	868	48	13
C	750	45	810	64	19
D	726	54	639	77	23
E	1480	73	434	86	13
O/N	1495	92	315	92	0
U	666	100	401	100	0
Total	7890		5011		

(Unfortunately disaggregated statistics for the 1986 syllabuses were not available at the time of writing.)

In table 3, the total grade distributions for mathematics syllabuses for all the boards are given (in 1986 some small syllabuses were not included in the inter-board statistics so the total entry is slightly under-estimated). The change over time is similar to that reported for the UCLES syllabuses only.

Table 3: Grade distribution for mathematics syllabuses offered by all boards - 1986-1995

Grade	1986 Syllabuses number	Cum %	1995 number	Cum %	Difference
A	6848	16	15384	26	10
B	6805	32	10694	44	12
C	4828	43	9911	61	18
D	5573	56	8498	76	20
E	6866	72	6382	87	15
O/N	6328	86	4054	93	7
U	5917	100	3853	100	0
Total	43163		58776		

There are three possible explanations for the grade distributions in Tables 2 and 3 for the increasing percentage achieving each grade:

- either the quality of entry from the UCLES syllabuses improved as a result of weaker candidates opting for other syllabuses or other subjects but there had been a overall decline in grading standards;
- or an overall decline in grading standards for all boards;
- or an improvement in the quality of candidates.

One way of investigating how the A-level entry has changed is to consider the effect of centre type. On average, candidates from different types of centres tend to perform differently. For example, 41.4% of candidates from independent schools obtained a grade A on the 1995 9205 UCLES syllabus compared with just 22.9% of candidates from comprehensive schools. The distributions by centre type for the UCLES linear syllabuses and for the mathematics syllabuses from all boards are given in Table 4.

Table 4: Entry of A-level mathematics by centre type for the UCLES syllabuses and for all boards
(Percentage of candidates)

Centre Type	UCLES 1986	UCLES 1995	All Boards 1986	All Boards 1995
Independent	21	33	15	20
Selective	12	16	12	12
Comprehensive	46	34	41	36
Sixth form Colleges	16	15	11	15
Others	5	2	21	18

(Others includes private candidates, and mainly, F.E. establishments)

From Table 4, there is evidence to support an increase in quality of candidates for the UCLES syllabus because the proportion of candidates from independent and selective schools has increased (in terms of numbers, the change is caused by comprehensive schools changing to other boards). For all boards, the entry distribution would suggest a smaller improvement. There have been a number of comparability studies between mathematics syllabuses for A-level boards which have usually found that the UCLES mathematics syllabuses tended to be one of the most, if not the most rigorous mathematics syllabus (Bell, Bramley and Raikes, in prep.). This has resulted in pressure to change the UCLES examinations in this period. The result of this change is described in the next section. The data presented in this section suggests that there is some justification for a positive changes in the UCLES A-level grade distribution.

Assessment of question difficulty

Nuttall (1979) recommended that panels of judges be used to grade standards. He noted two advantages to using panels of expert judges to assess changes in standards. This first advantage is that they can declare the comparison invalid because the syllabuses have changed so much as to make direct comparison impossible. Fortunately this was not a problem in this case but Christie and Forrest (1980) found that it was the case in a comparison of Chemistry over the period 1963 to 1973. The second advantage is that the panel can make subtle adjustments for change in the content or coverage of the test and assign less weight to skills and topics that are declining in importance, and more to those that are increasing in importance. In this study, the judges were asked to complete two tasks. Firstly, they were asked to assess the difficulty of the individual questions. Then they were asked to compare scripts using Thurstone paired comparison methodology.

The work described in the folowing section section used two panels of scrutineers. These scrutineers worked on the UODLE and OCSEB A-level examinations and not recent UCLES examinations (UODLE and OCSEB are now part of UCLES but to honour commitments to existing candidates separate A-level examinations have been continued). Although the scrutineers were not told the age of the scripts they were told that they should assume that the candidates had been prepared for the examination in question. The scrutineers were not told that the study was about comparisons over time. It was described as an investigation into the use of Thurstone Paired Comparison Methodology as a method of awarding when there was a change of syllabuses.

The scrutineers were also asked to assess the difficulty of the individual questions in the examination papers using the following scale:

1 Much less demand than a typical A-level question

2 Slightly less demand than a typical A-level question

3 The demand of a typical A-level question

4 Slightly more demand than a typical A-level question

5 Much more demand than a typical A-level question.

These ratings can be used to investigate changes in the perceived accessibility of the papers. The scrutineers were sent the papers and mark schemes before the study and were given a series of questionnaires (one for each paper) to complete before the residential meeting. The main reason for this exercise was to help the scrutineers familiarise themselves with the examination papers before the study.

These ratings were analysed using a generalised linear models procedure that corrects for the missing values. The resulting ANOVA table is presented below:

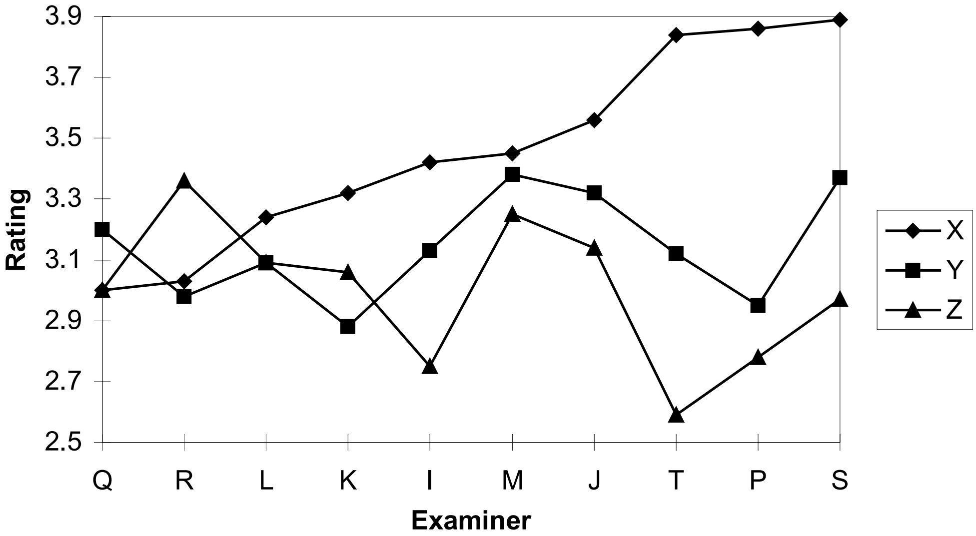
Source	DF	Type I SS	Mean Square	F Value	Pr> F
Rater	9	14.47	1.61	3.16	0.0009
Syllabus	2	35.44	17.72	34.81	0.0001
Syllabus *rater	18	43.71	2.43	4.77	0.0001

This indicates that the ratings, syllabus and the interaction are all significant.

To investigate this pattern further the least square means were calculated. Least square means are used because of the missing ratings in the data set. The results are given in Figure 1 below.

All the raters except Q and R considered the X syllabus to be the most demanding. In addition, examiner L considers the Y syllabus to have the same demand as syllabus X. Overall most of these expert examiners have concluded that the questions have become less demanding since 1986.

Figure 1: Least mean square rating by examination and syllabus
(X - 1986 9200, Y - 1986 9205, Z - 1995 9205)



Thurstone Paired Comparison Methodology

The scrutineers were formed in two teams (A and B). It was intended that Team A consisted of 3 UODLE examiners and 2 OCSEB examiners and team B 2 UODLE examiners and 3 OCSEB examiners. However, one of the OCSEB examiners had to withdraw at the last minute because of illness so team B consisted of 2 UODLE and 2 OCSEB examiners.

The two teams were asked to make many series of comparisons of pairs of scripts. For each comparison they were required to nominate one of the pair as the better. The judges were required to make a forced choice and not allowed to have ties. With Thurstone's method, equality between two scripts A and B is usually expressed by half of the judges voting A better and half voting B better (although, in this study, there is an odd number of judges so there cannot be ties).

With this methodology, a judge is asked to compare the quality of one script with another and not to an internalised mental standard for a level. This methodology has two advantages. Firstly, concrete comparisons between two scripts are made, removing the uncertainties associated with a notional standard. Secondly, differences between notional standards of judges cancel out and the methodology, therefore, controls for variability in judges' internal standards.

It was decided that scrutineers should use copies of the scripts rather than the originals. There were two reasons for this. Firstly, it allowed the concealing of the total marks and details of the syllabuses from the scrutineers. Secondly, the scrutineers would not be able to identify the age of the scripts (even if they recognised the old syllabuses they would not know whether the data came from modern candidates or original candidates). The ticks and other marks made by the original examiners were written over using appropriately coloured highlighter pens. Although the total marks were concealed, it was impossible to prevent the examiners from adding the marks up again if they wished.

For the 1995 syllabus, ten scripts on the boundary mark were selected at the A, B and E boundaries. At the E boundary some scripts on either side of the boundary mark had to be selected because of the relatively small number of scripts at this part of the distribution). For the 1986 syllabuses, seven scripts were chosen in an attempt to satisfy the following pattern:

boundary + 3x marks
boundary + 2x marks
boundary + 1x marks
boundary
boundary - x marks
boundary - 2x marks
boundary - 3x marks

where x was approximately 1/3 of the A/B range for the A and B boundaries and the D/E range for the E boundary. Because there was an overlap for the sets for the A and B boundaries, different scripts were used for the two boundary comparisons. In addition, because of the small number of scripts, scripts with a mark that exactly satisfied the above criteria were not always available and the script nearest the mark was chosen. Only five of the seven scripts were to be used in the study. Usually scripts from the range +2 to -2 were used in study but for some boundaries the range was offset by a third of a grade. This was to demonstrate that the examiners were not basing their decisions on matching the 1995 to the middle script of the 1986 set (obviously they would only be able to do this after the first cycle of comparisons). The unused extreme scripts had been prepared for use on the second day if the first day produced evidence of a substantial change in grading standards.

The 1/3 of grade range was chosen for two reasons. Firstly, there is a limitation on the accuracy of awarding. There are bound to be small differences between adjacent years. Secondly, it was decided that the range chosen should be large enough to detect substantial changes in grading standards. The scripts were chosen so that both mechanics and statistics questions were represented at each boundary. This choice was restricted by the relatively small number of old scripts that was available. It should be noted that the rank order of scripts for the individual papers differs from that of the total mark. Because of the small number of scripts it was not possible to choose scripts with even profiles of component paper marks. The implications of this are discussed in the result sections.

On the first day, team A compared 1986 9200 scripts with 1995 scripts and team B compared 1986 9205 scripts with another set of 1995 scripts. On the second day, the teams swapped 1986 syllabuses.

At the start of this study the examiners were told that the purpose of this study was to test a new methodology for maintaining standards when syllabuses were changed. This was not untrue because the operational results of this study are being considered by a committee that is addressing this issue. They were not told the identity or the age of the syllabuses, or the age of the scripts. Given that the scrutineers were senior A-level examining personnel, it was clear that they had a some idea of the age of the syllabuses. There are some clues to the age of the examinations papers in the wording of some of the questions, e.g., the price of stamps. Indeed one scrutineer who was a teacher used the 9205 syllabus in his school. They did seem to have one misconception. Some of the scrutineers thought that the 9200 examination was at least fifteen years-old rather than ten years-old (This is not so surprising since the

9200 syllabus was in the process of being superceded). However, for a scrutineer to generate data indicating no change by using the total mark instead of making judgements, he would have had to have known what the grade boundaries were for the 1986 syllabuses. The reason these precautions were taken was not because the scrutineers were not trusted but to defuse any potential criticism of the results.

The design for any individual boundary involved examiners passing pairs of scripts cyclically round a table and at the end of each cycle reconstituting the pairs of scripts. Full details of the design will be given in the main report (Bell, Bramley and Raikes, in prep.). On the first day the results were monitored to check that the range of scripts covered the appropriate ranges to detect the magnitude of the changes. Since there was no evidence that they were not, the scripts were not changed.

The results presented in this paper are preliminary and are based on the simplest summary statistics. The first set of comparisons relates to the syllabus 1986 9200. This syllabus was the oldest in the study and the style of questions had given some scrutineers the impression that it was older than it was. It should also be noted that this syllabus differed most from the 1995 syllabus which obviously added an extra level of difficulty to the comparisons.

Results for 9200 1986 vs 9205 1995

At the Grade A boundary (see Table 5), script XA2 (i.e., a script from the 1986 9200 syllabus) was judged better than a script on the grade A boundary for the 1995 9205 syllabus 39 out of 45 times. In the table the total marks for paper 1 and paper 2 and the overall total mark are also given. When two scripts are of equal quality then they should each win roughly 50% of the time. From these summary statistics, it is clear that XA5, a grade B script, has been judged to be better than the 1995 grade A scripts. The dotted line indicates where the percentage of wins changes from greater than 50% to less than 50%. Scripts XA1 and XA4 have been placed in the wrong order according to the total mark but not in respect of the paper 1 mark. Because of the length of the scripts and the greater similarity between papers 1 and paper 2 it appears that the quality of the paper 1 response may have given more weight by scrutineers than 2.

Not all the comparisons were completed at the grade B boundary. A 1986 grade C script XB3 was considered better than 1995 borderline B/C scripts. This suggests a small decline in standards. At the grade E boundary, the range of scripts on this boundary had been offset so that the worse script was approximately the whole of the D/E range from the boundary mark. There was no evidence of a change in standards.

Table 5: Results for 9200 1986 vs 9205 1995

Grade A Boundary

Script	Paper 1	Paper 2	Total mark	Grade	Wins	Losses	Total	% Win
XA2	88	77	165	A	39	6	45	87
XA3	88	71	159	A	36	9	45	80
XA5	83	61	146	B	31	14	45	69
XA1	80	57	137	B	13	32	45	29
XA4	68	74	142	B	12	33	45	27

Grade B Boundary

Script	Paper 1	Paper 2	Total Mark	Grade	Wins	Losses	Total	%
XB2	67	64	131	B	25	3	28	89
XB5	76	64	140	B	23	5	28	82
XB1	58	68	126	B	20	8	28	71
XB3	69	52	121	C	16	12	28	57
XB4	47	70	117	C	7	21	28	25

Grade E Boundary

Script	Paper 1	Paper 2	Total mark	Grade	Wins	Losses	Total	%
XE1	43	45	88	E	40	5	45	89
XE5	45	37	82	E	36	9	45	80
XE4	37	33	70	O	20	25	45	44
XE2	25	39	64	O	17	28	45	38
XE3	31	27	58	O	12	33	45	27

Results for 9205 1986 vs 9205 1995

There is evidence of *no change* in grading standards at the grade A and B boundaries for the comparison of the 9205 syllabuses but some evidence that there was an *increase* in standards at grade E. The lowest grade A script used in 1985 was almost exactly equivalent to the grade A scripts from 1995 because it won nearly half the comparisons. The improved accessibility of the 1995 paper meant that candidates at grade E were more able to demonstrate the extent of their knowledge which could aid the grading at the grade E boundary. These results demonstrate that changes in accessibility to do not necessarily lead to a change in the grading standards.

Table 6: Results for 9205 1986 vs 9205 1995

Grade A Boundary

Script	Paper 1	Paper 2	Total Mark	Grade	Wins	Losses	Total	%
YA4	84	82	166	A	37	8	45	82
YA5	87	71	158	A	21	24	45	47
YA1	70	68	138	B	19	26	45	42
YA3	81	66	147	B	17	28	45	38
YA2	68	62	130	B	7	38	45	16

Grade B Boundary

script	Paper 1	Paper 2	Total Mark	Grade	Wins	Losses	Total	%
YB3	72	64	136	B	33	12	45	73
YB4	78	66	144	B	30	15	45	67
YB5	70	58	128	B	27	18	45	60
YB2	76	44	120	C	8	37	45	18
YB1	74	43	117	C	4	41	45	9

Grade E Boundary

Script	Paper 1	Paper 2	Total Mark	Grade	Wins	Losses	Total	%
YE2	44	48	92	E	39	6	45	87
YE5	47	40	87	E	29	16	45	64
YE1	47	32	79	E	16	29	45	36
YE3	56	15	71	O	10	35	45	22
YE4	46	20	66	O	8	37	45	18

A more detailed analysis of these results will appear in the main report (Bell, Bramley and Raikes, in prep.) which is in preparation.

Discussion

In this paper, it has been shown that there has been an increase in the percentages of candidates obtaining high grades for the UCLES linear A-level mathematics syllabuses. Two questions have to be considered. Firstly, is the change the result of decline in the grading standards or an improvement in the quality of the candidates or a combination of both? Secondly, how important is this change?

The results of the Thurstone comparisons show that although there was a decline in a grading standards between the 1986 9200 syllabus and the 1995 9205 syllabus, for the 1986 9205 syllabus and the 1995 9205 syllabus, there was *no change* for the grading standards for the A and B boundaries and a *small increase* in grading standard at the grade E boundary. Given the status of the 9200 syllabus, this would suggest that if there has been a decline in standards it occurred in the mid-eighties. Because of the attempts to make the examination more accessible, there is more evidence of mathematics ability available in the scripts at the grade E boundary. It should be recognised that ongoing work on the Thurstone study data using more sophisticated analyses suggest it is not capable of detecting small changes in standards.

This mean that the improvement in the grade distribution must be explained by an improvement in the candidates. The evidence for such a change is complex. For the UCLES linear A-level mathematics syllabuses the entry has declined in numbers but the distribution by centre type suggests an improvement in the quality of the entry. This argument does not apply in general. The change in cohort size would suggest that quality of entry has declined unless in 1986 there was a substantial number of candidates who could have obtained high grades in A-level mathematics but did not continue their studies after sixteen. However, even if there were such candidates capable of taking A-level mathematics in 1986 it would not be unreasonable to assume that the effect of increased uptake would have been greater for lower grades than higher. The changes prior to the start of sixth form studies also suggest that the quality of entry for A-level mathematics has not improved. However, assuming that higher education has maintained standards, then there is support for an improvement in the quality of A-level candidates.

By combining all the evidence in this paper, it seems reasonable to conclude that the improvement in the grade distribution for the UCLES syllabuses only is a mixture of the following: the withdrawal of the difficult 9200 syllabus which suggests a change of standard in mathematics in the early 1980's, an improvement in the quality of the entry, and possibly a small change in grading standards within the tolerances of what can be achieved by current awarding methods and the Thurstone paired comparison methodology.

The second question is much more complex. The content of the A-level mathematics curriculum and what a particular grade should indicate about the mathematical ability of a candidate should be the subject of consultation and debate. It is, however, worth noting that there is considerable evidence that the standard of mathematics attained by English eighteen-year-olds who specialise in mathematics is higher than that in other countries. For example, Prais (1986) noted that

.. comparative tests in the 1960s showed that those English school-leavers specialising in mathematics at A-level reached standards significantly higher than in Germany and, indeed, not exceeded in any other advanced country at schools catering for those of academic ability. Preliminary results of similar comparisons for 1981 indicate that England's advantage for this special group has been maintained.'

Some authors have questioned whether this highly selective and demanding examination system is the best policy. It can be argued that it is better have a reasonable number of good mathematicians at 18 rather than a small but excellent elite. For example, Prais (1986) reported the findings of a British engineer who observed standards of engineering courses in five Japanese universities. The engineer concluded that Japanese engineering graduates although not as adventurous in their thinking as British graduates reached a standard which enables them to put into routine production the most advanced of production methods, and to make suitable use of foreign patents. Eckstein and Noah (1993) noted that in 1987 the Japanese examination system established a somewhat lower level of difficulty in its mathematics examinations for 17- and 18-year-olds. This did ensure that a higher proportion of the age group took mathematics examinations than in England or Germany.

The results of the analysis of examiners' ratings and the inspection of the scripts indicate that there has been an attempt to make the examination more accessible over the period. Reid (1991) argued that an 'accessible' curriculum need not fall into the trap of dilution and diffusion. He cited the case of British Columbia where high enrolment has been combined with good levels of attainment to yield higher overall productivity (McKnight, 1987).

References

Bell, J.F., Bramley, T., and Raikes, N. (In prep.) *Standards in A-level Mathematics. 1986-1995.* Cambridge UCLES.

DES (1994a) Statistics of schools in England - January 1993. *Statistical Bulletin*, 8/94.

DES (1994b) GCSE and GCE A/AS Examination results 1992/93. *Statistical Bulletin*, 7/94.

Eckstein, M.A., and Noah, J. (1993*) Secondary School Examinations. International Perspectives on policies and practices*. New Haven and London: Yale University Press.

Fitz-Gibbon, C.T., and Vincent, L. (1994) *Candidates' Performance in Public Examinations in mathematics and Science.* London: SCAA.

Foxman, D., Ruddock, G., McCallum, I., and Schagen, I. (1991) *APU Mathematics Monitoring (Phase 2)*. Slough: National Foundation for Educational Research.

Goldsmith, C. (1997) Old standard fails to pay. *Times Educational Supplement*, No. 4207, 14 February.

HEQC (1995) *Graduate standards programme: Interim report.* London: Higher Educational Quality Council.

HESA (1995) *Higher education statistics for the United Kingdom 1992/93.* Cheltenham: Higher Education Statistics Agency.

London Mathematical Society, Institute of Mathematics and its Applications and Royal Statistical Society (1995) *Tackling the Mathematics Problem.* London: London Mathematical Society.

Kahn, P., and Hoyles, C. (1997). The changing undergraduate experience: a case study of single honours mathematics in England and Wales. *Studies in Higher Education,* 22, In press.

McKnight, C.C. (1987) The underachieving curriculum: assessing US mathematics from an international perspective. Champaign, Illinois: Stipes Publishing.

Nuttall, D.L. (1986) Problems in the measurement of change. In Nuttall, D.L. (Ed) *Assessing Educational Achievement*. Lewes: Falmer Press.

OFSTED (1996) *GCE Advanced Supplementary and Advanced Level Examinations 1996. A report from the Office of Her Majesty's Chief Inspector of Schools.* London: The Stationery Office.

OFSTED/SCAA (1996) *Standards in Public Examinations 1975 to 1995*. London: OFSTED/SCAA.

Prais, S.J. (1986) Education for productivity: comparisons of Japanese and English schooling and vocational preparation. *Compare*, 16, 121-147.

Reid, W. (1991). The ideology of access in comparative perspective. In C. Chitty (ed.), *Post-16 Education. Studies in Access and Achievement*. London: Kogan Page.

Smithers, A., and Robinson, P. (1996) *Trends in Higher Education*. London: The Council for Industry and Higher Education.

Sutherland, R., and Pozzi, S. (1995) *The Changing Mathematical Background of Undergraduate Engineers.* London: Engineering Council.

Wiliam, D., Brown, M., Kerslake, D., Martin, S., and Neil, H. (1996) *Is the gap between GCSE and A-level bigger in Mathematics than for other subjects?* Paper presented at the 22nd Annual British Educational Research Association Conference, University of Lancaster, September 1996.

Ziomek, R.L., and Svec, J.C. (1995) *High School Grades and Achievement: Evidence of Grade Inflation*. Iowa: ACT.