Investigating A level mathematics standards over time

Every year there has been a debate about the increasing proportion of candidates obtaining A levels. Mathematics has been a particular area of concern. It has been argued that this is not because the candidates are better, but because the examination is 'easier'. The London Mathematical Society, the Institute of Mathematics and its Applications and the Royal Statistical Society (LMS/IMA/RSS 1995) suggested that new higher education students were hampered by a lack of fluency and reliability in numerical and algebraic manipulation and simplification, and that there had been a decline in students' analytical powers when faced with a multi-step problem. This article reports the results of a study of 1985 and 1996 A level mathematics scripts using Thurstone comparison methodology.

John F. Bell, Tom Bramley *and* **Nicholas Raikes,**
University of Cambridge Local Examinations Syndicate

Because the consensus of opinion about the relative importance of component skills and knowledge within a subject area changes over time, these changes have to be reflected in A level syllabuses and examination papers. This means that differences between performance and expectations must be considered. For example, in an 1863 examination, candidates were asked the following:

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

Most modern A level mathematics candidates would not be able to calculate that the answer was £95 17s. 6¾d. because of their unfamiliarity with the units. While there almost certainly has been a decline in ability to perform calculations using such units, this would not be considered important. 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.

To discuss changes over time, it is necessary to define different types of changes. These are:

- **change in coverage**, i.e. the content of the syllabus and the range and relative importance of the material covered;
- **change in accessibility**, i.e. the relative difficulty of obtaining correct answers for the questions in an examination;
- **change in grading standard**, i.e. the difficulty of obtaining a particular grade.

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 an overall change in accessibility. 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, candidates may spend less time studying it and understanding may decline. If this material is vital for a course of higher education, then the lecturers of such candidates will conclude that there has been a decline in standards. The fact that other parts of the syllabus have been given greater emphasis may not be noticed or appreciated.

There are two reasons why a grade distribution for an examination has improved: the *grading standard* has changed and/or the *quality of candidates* has changed. Not all changes in the quality of the entry result from a general improvement in candidate quality. At syllabus level, candidate quality can improve when weaker candidates move to another syllabus. At 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.

COMPARISON OF SYLLABUSES

Three UCLES syllabuses were considered in this study: Mathematics Syllabus A 9200 in 1986, and Mathematics Syllabus 9205 in 1986 and 1995. 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 there had been no change in the 9205 syllabus, there had been changes in the style and layout of the Paper 2. In 1995 the questions tended to have more structure than those in 1986. However, the *pure mathematics* questions had changed the least over

the nine years and appeared to be similar in structure and content. 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 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. For the *statistics* questions, the range of statistics covered was identical and the style of the questions was very similar. The 1995 paper seemed more varied and interesting, with more attempt to give the questions a relevant context (e.g. vegetable crops) rather than a very artificial context (e.g. boxes of red and green apples; random arrivals of broken noses at a surgery). Unlike the mechanics questions, both papers contained questions that would require a calculator.

The grade distributions for UCLES linear mathematics syllabuses are given in Table 1. The entries for the UCLES linear mathematics syllabuses had declined by 2869 candidates (36%). Given the cohort size had also declined over the period, a decline in entry would be expected but for the fact that the overall entry for mathematics for all boards had increased. There had also been a large increase in the cumulative percentages of candidates – for example, the percentage of candidates obtaining at least a grade E rose from 72.6% in 1986 to 85.7% in 1995.

In Table 2, 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.

There are three possible explanations for changes in the grade distributions in Tables 1 and 2:

- an overall decline in grading standards for all boards;
- an overall improvement in the quality of candidates;
- 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 an overall decline in grading standards.

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 22.9% of candidates from comprehensive schools. The distributions by centre type are given in Table 3. The quality of candidates for the UCLES syllabuses has improved 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 has been a number of comparability studies

Table 1: Grade distributions for UCLES linear mathematics syllabuses

Grade	1986 UCLES Syllabuses		1995 UCLES *9205*		
	Number	*Cum %*	*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		

Table 2: Grade distributions for mathematics syllabuses offered by all boards, 1986–1995

Grade	1986 Syllabuses		1995		
	Number	*Cum %*	*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		

Table 3: Entry of A level mathematics by centre type for 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

*FE establishments and private candidates

between mathematics syllabuses for A level boards which have usually found that the UCLES mathematics syllabus tended to be one of the most, if not *the* most, rigorous (Bell, Bramley and Raikes, in preparation).

THURSTONE COMPARISON METHODOLOGY

Nuttall (1986) recommended that panels of judges be used to compare standards. He noted two advantages for this method. First, 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. Second, 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 vice versa. In this study, the judges used Thurstone paired-comparison methodology.

This study used two teams 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). 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 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 an exact tie). 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. First, concrete comparisons between two scripts are made, removing the uncertainties associated with a notional standard. Second, differences between notional standards of judges cancel out and the methodology, therefore, controls for variability in judges' internal standards.

For the 1995 syllabus, ten scripts on the boundary mark were selected at the A, B and E boundaries. 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 ⅓ of the A/B range for the A and B boundaries and the D/E range for the E boundary. Only five of the seven scripts were to be used. Usually scripts from the range +2 to –2 were used in the 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 scripts 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.

On the first day, team A compared 1986 *9200* scripts with 1995 *9205* scripts, and team B compared 1986 *9205* scripts with another set of 1995 *9205* 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 really was a secondary aim of the study). Given that the scrutineers were senior A level examining personnel, it was clear that they had some idea of the age of the syllabuses. They did seem to have one misconception. Some of the scrutineers thought that the *9200* examination was at least fifteen years rather than ten years old. (This is not so surprising, since the *9200* syllabus was obsolescent in 1986.) However, for a scrutineer to generate data indicating no change by using the total mark, he would have had to have known what the grade boundaries were for the 1986 syllabuses. 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. On the first day the results were monitored to check ranges were appropriate, but there was no need for any changes. Further details of the study can be found in Bell, Bramley and Raikes (in preparation).

Table 4: Results for 9200 1986 vs 9205 1995

Grade A Boundary						
Script	**Total mark**	**Grade**	**Wins**	**Losses**	**Total**	**% win**
XA2	165	A	39	6	45	87
XA3	159	A	36	9	45	80
XA5	146	B	31	14	45	69
XA1	137	B	13	32	45	29
XA4	142	B	12	33	45	27
Grade B Boundary						
Script	**Total mark**	**Grade**	**Wins**	**Losses**	**Total**	**% win**
XB2	131	B	25	3	28	89
XB5	140	B	23	5	28	82
XB1	126	B	20	8	28	71
XB3	121	C	16	12	28	57
XB4	117	C	7	21	28	25
Grade E Boundary						
Script	**Total mark**	**Grade**	**Wins**	**Losses**	**Total**	**% win**
XE1	88	E	40	5	45	89
XE5	82	E	36	9	45	80
XE4	70	O	20	25	45	44
XE2	64	O	17	28	45	38
XE3	58	O	12	33	45	27

Results for *9200* 1986 vs *9205* 1995

For this particular comparison, there is evidence of a *decline* in standards at the grade A and B boundaries and no change at the grade E boundary. At the grade A boundary (see Table 4), 1986 *9200* script XA2 was judged better than an A boundary 1995 *9205* script 39 out of 45 times. 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%. Because of time pressure, not all the comparisons were made at the grade B boundary. Again there is evidence for a small decline in standards. At the grade E boundary, the range of scripts on this boundary had been offset so that the worst script was approximately the whole of the D/E range from the boundary mark. There was no evidence of a change in standards at the E/O boundary.

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 (as indicated by the dotted lines). The improved accessibility of the 1995 paper meant that candidates at grade E were better 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 do not necessarily lead to a change in grading standards.

DISCUSSION

Because 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. First, is the change the result of a decline in the grading standards or an improvement in the quality of the candidates, or a combination of both? Second, how important is this change?

The results of the Thurstone comparisons show that although there was a decline in grading standards between the obsolescent 1986 *9200* syllabus and the 1995 *9205* syllabus, for the 1986 *9205* syllabus and the 1995 *9205* syllabus, there was *no change* in 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, if there has been a decline in standards it occurred in the mid-eighties. Because of attempts to make the examination more accessible, there is more evidence of mathematics ability available in the scripts at the 1995 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.

There is evidence of an improvement in the candidates for the UCLES linear A level mathematics because of the change in distribution by centre type. For the UCLES syllabuses, the cause of the change in grade distribution is a

Table 5: Results for 9205 *1986 vs* 9205 *1995*

Grade A Boundary						
Script	**Total mark**	**Grade**	**Wins**	**Losses**	**Total**	**%**
YA4	166	A	37	8	45	82
YA5	158	A	21	24	45	47
YA1	138	B	19	26	45	42
YA3	147	B	17	28	45	38
YA2	130	B	7	38	45	16
Grade B Boundary						
Script	**Total mark**	**Grade**	**Wins**	**Losses**	**Total**	**%**
YB3	136	B	33	12	45	73
YB4	144	B	30	15	45	67
YB5	128	B	27	18	45	60
YB2	120	C	8	37	45	18
YB1	117	C	4	41	45	9
Grade E Boundary						
Script	**Total mark**	**Grade**	**Wins**	**Losses**	**Total**	**%**
YE2	92	E	39	6	45	87
YE5	87	E	29	16	45	64
YE1	79	E	16	29	45	36
YE3	71	O	10	35	45	22
YE4	66	O	8	37	45	18

mixture of the withdrawal of the difficult and obsolete *9200* syllabus and 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 detected by 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 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 (Prais 1986). However, it can be argued that it is better to 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 Japanese engineering courses. The engineer concluded that Japanese engineering graduates, although not as adventurous in their thinking as British graduates, reached a standard which enabled 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 lower level of difficulty in the Japanese mathematics examinations ensured that a higher proportion of the 17–18 year old year-group took them compared with England or Germany. 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. (1997) *Standards in A level Mathematics, 1986-1996*, paper presented at the British Educational Research Association, University of York.

BELL, J.F., BRAMLEY, T. and RAIKES, N. (in preparation) *Standards in A level Mathematics, 1986-1995*, University of Cambridge Local Examinations Syndicate.

ECKSTEIN, M.A. and NOAH, J. (1993) *Secondary School Examinations. International Perspectives on policies and practices*, Yale University Press.

LONDON MATHEMATICAL SOCIETY, INSTITUTE OF MATHEMATICS AND ITS APPLICATIONS AND ROYAL STATISTICAL SOCIETY (1995) *Tackling the Mathematics Problem*, London Mathematical Society.

NUTTALL, D.L. (1986) 'Problems in the measurement of change', in Nuttall, D. L. (ed), *Assessing Educational Achievement*, Falmer Press.

PRAIS, S.J. (1986) 'Education for productivity: comparisons of Japanese and English schooling and vocational preparation', *Compare, 16*, 121-47.

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