Changes in standards at GCSE and A-Level: Evidence from ALIS and YELLIS

Report for the ONS by Robert Coe

CEM Centre, Durham University

April 2007

This report is an update to part of a report written for the Sunday Times in April 2005[1]. It presents the evidence from data collected for schools to monitor their own performance as part of the YELLIS and ALIS projects.

About ALIS and YELLIS

ALIS (the A Level Information System) began in 1983 as a system for helping schools to compare the progress their students have made with that of students in other schools. Currently over 1400 schools and colleges participate in the project, which processes about half of the A levels taken in the UK. Schools receive value added analysis for individual subject entries, based on simple residual gains when A level scores are regressed on average GCSE scores, as well as data on a range of student attitudes and perceptions. An optional part of the scheme is the Test of Developed Abilities (TDA), which is offered free of charge to participants in ALIS, should they wish to have an additional base-line measure from which to calculate value added.

YELLIS (Year 11 Information System) began in 1994 and now analyses the GCSE results of about 1300 schools. YELLIS uses its own test, taken by students in year 10 or year 11, as a base-line from which to calculate value added. The YELLIS test comprises two sections, mathematics and vocabulary.

For the purposes of this analysis, both projects provide two kinds of data. The first is a set of scores on a measure of general academic ability that has remained constant over a number of years. This avoids one of the problems of trying to compare GCSE or GCE grades over time which is that the examination is different each year. Inevitably, though, it raises further problems of whether the test is appropriate to use as a measure of attainment and, if so, whether its appropriateness remains constant over the time period in question. These are controversial matters and the subject of dispute by academics and examiners.

The second is data on the relationship between those ability test scores and subsequent performance in national examinations. Knowing the ability of students

[1] The original report can be found at http://www.times-archive.co.uk/onlinespecials/english_in_schools.html

who have taken a particular exam in a particular year enables us to compare their achievement with other students of similar ability in other years.

Achievement at KS4: Evidence from YELLIS

Changes in performance on the YELLIS test

The Yellis test average scores for all students who took the test in either Year 10 or Year 11 between 1995 and 2006 are shown in Figure 1. Given that the numbers are quite small in the early years of Yellis, we should interpret the results for 1995 and 1996 with caution. The numbers taking the Y11 test in 2005 and 2006 are also relatively small (under 4000), so the dip in scores for those years may be less reliable.

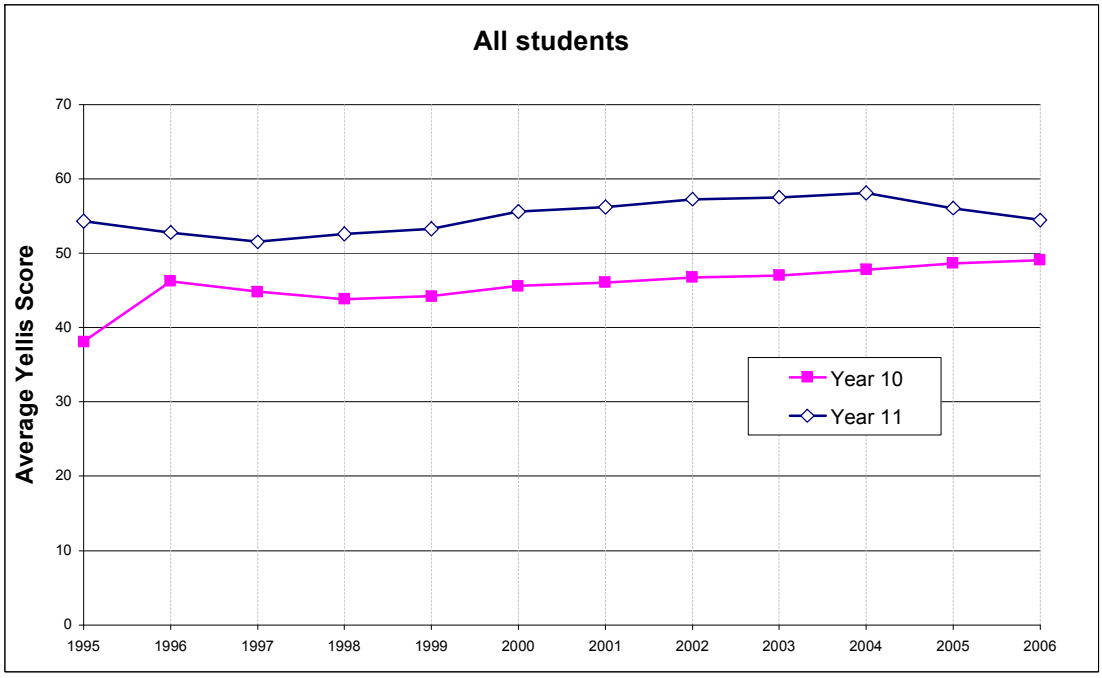

Figure 1: Changes in performance on the YELLIS test over time

There appears to have been a small but steady rise in the average scores of Y10 students. Between 1998 and 2006 the increase amounts to an effect size of 0.3[2]. The fact that the scores are increasing suggests that students' mathematical and verbal abilities may have improved slightly over that period, though the increase on this scale over eight years is quite small. Nevertheless, this is certainly consistent with the claim that standards of educational attainment have genuinely risen.

Of course, any interpretation of this change depends on the assumption that both groups of students are representative of their respective year groups. Some evidence to support this assumption comes from a study by Telhaj et al (2004) who considered a sample of 664 schools with five continuous years of YELLIS

[2] i.e. 0.3 of the population standard deviation for the test.

membership that included the year 1998. They found that this group was very close to being representative of the national population of schools on almost all variables available.

Changes at GCSE in relation to the Yellis test

Against the background of substantial continuing rises in national GCSE performance and a slight but steady rise in Yellis test scores (Figure 1), the question remains whether students with the same Yellis score achieve more or less at GCSE over time.

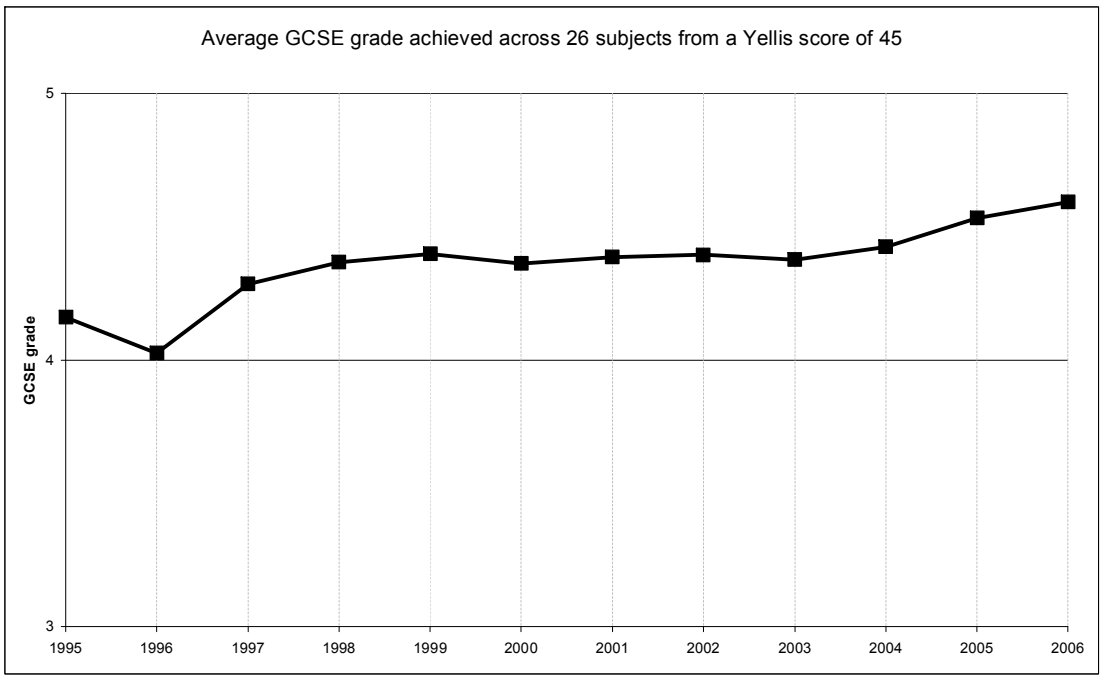

Figure 2: Change in average GCSE grade, controlling for ability

For this analysis, we take the relationship between the Yellis test taken in Y10 and GCSE grades achieved in examinations about 20 months later. The numbers of students in this group are much larger than for the Y11 group for all but the first few years of Yellis. For simplicity we take a typical Yellis test score of 45 as our reference point. The comparison therefore shows the average grade achieved by students who scored 45 on the test in Y10, and hence controls for the ability of the student.

Overall, the averages of all GCSE grades achieved each year by these students with 45 on the Yellis test are shown in Figure 2. Twenty-six separate GCSE subjects were analysed by Yellis throughout this period. GCSE grades are coded numerically as A*=8, A=7, B=6, C=5, D=4, E=3, F=2, G=1, U=0. Again if we discount 1995 where the numbers are smaller (5000 students) and the membership of Yellis less stable, there appear to be three phases.

Between 1996 and 1998 we see an apparent increase. Over two years, average performance increases by about a third of a GCSE grade per subject, controlling for ability.

For the period 1998 – 2003, performance seems fairly stable, so for all practical purposes, the line is flat. For GCSE subjects as a whole, students of comparable ability achieved the same grades regardless of which year they took their exams during this period.

However, the latest data allow us to see the beginnings of an upwards trend from 2004 onwards. Between 2003 and 2006, candidates with the same ability have improved their performance by about a fifth of a GCSE grade on average per subject.

If we look separately at the five subjects with the largest number of entries (double science, English, French, history and maths), we see a somewhat mixed picture (Figure 3).

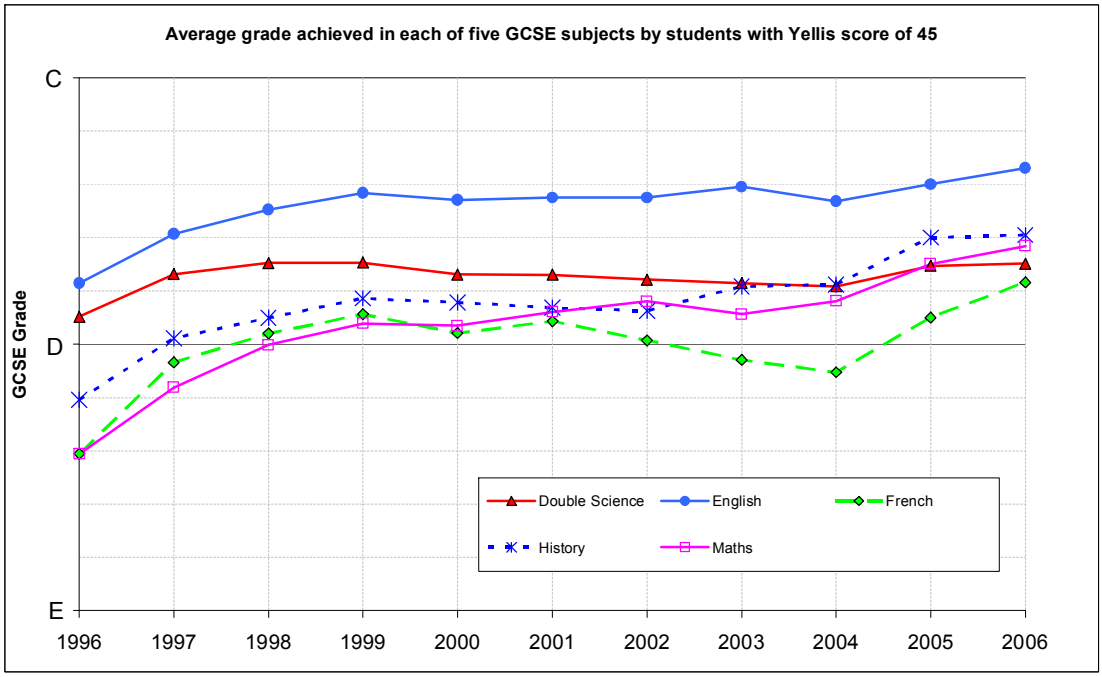

Figure 3: Change in GCSE grade achieved by students with the same ability scores (YELLIS test score =45)

Performance in all these subjects has increased over the period shown. Maths has increased the most at almost 0.8 of a grade. History and French have both risen about two-thirds of a grade overall, though the latter actually declined between 2001 and 2004, then rising steeply to catch up. English has risen fairly steadily, just under half a grade over the whole period. For science the rise has been smallest at only one fifth of a grade overall, and most of this in the first year (1996-7).

Changes at A-Level: Evidence from ALIS

Data showing the relationship between TDA scores and performance in a range of A levels are available from 1988. The six subjects with the largest entries in ALIS are analysed here: Biology, English (Literature), French, Geography, History and Mathematics.

There are some problems with making comparisons over such a period. Inevitably syllabuses change and it is not always straightforward to decide whether what is being compared is quite the same. In particular, a subject like mathematics includes a number of different syllabuses but excludes others. For example, until 2001 modular syllabuses were not included under this heading; in 2002, with the start of *Curriculum 2000*, all A level syllabuses effectively became modular. English is also somewhat problematic, with different syllabuses in literature and language, or mixtures of the two.

A further problem is that the TDA, used since 1988, was modified slightly in 2000 in order to improve its predictions. The original test was known as the International Test of Developed Abilities and the new test was known more simply as the Test of Developed Abilities (TDA). This change has resulted in a discontinuity between ITDA scores for students taking A levels up to 2001 and TDA scores for those taking them from 2002 onwards. Although the new test is slightly easier than the old one, in the analyses presented here a correction has been applied to equate scores from the two tests.

Changes in performance on the Test of Developed Abilities (TDA)

Changes in the TDA scores of candidates in the six subjects are shown in Figure 4.

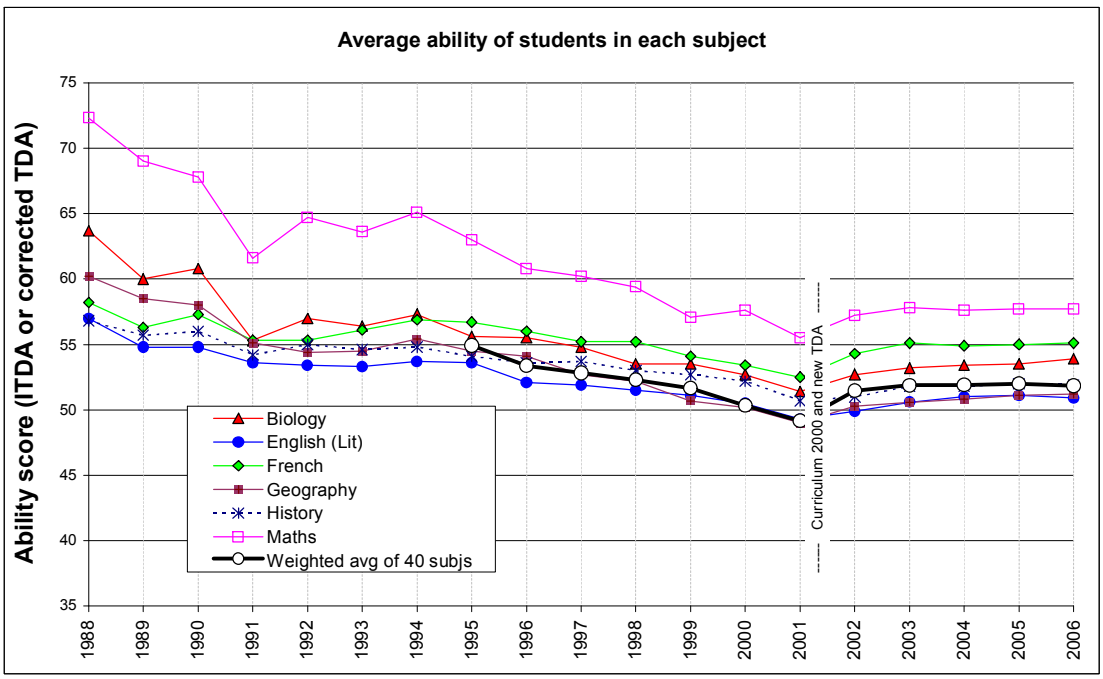

Figure 4 Mean ability score for students taking a range of A level subjects.

The data show that there has been a decline in the TDA scores of the candidates from 1988 to 2001, a jump between 2001 and 2002 and flat thereafter. The

coincidence of the 2001/2 jump with the change in the TDA/ITDA test might suggest that we need to be cautious about comparisons across the two versions of the tests. However, the change to modular syllabuses in 2002, where students who are thought unlikely to pass may well not actually enter the final examination, could also account for this apparent discontinuity.

A level results and TDA scores

Figure 5 shows the A level grades expected of candidates having the same TDA score each year since 1988 for the six subjects. From 1988 until 2006 the achievement levels have risen by about an average of 2 grades in each subject. Exceptionally, from 1988 the rise appears to be about 3.5 grades for Mathematics.

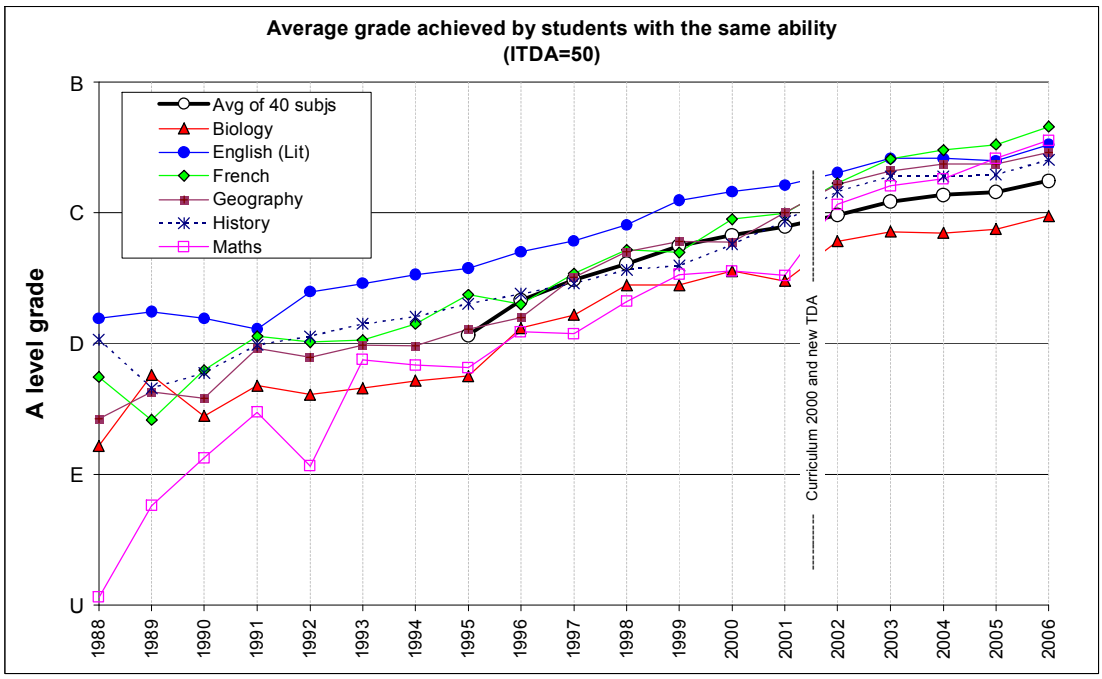

Figure 5: Mean A level grades achieved by students with the same ability scores (ITDA score of 50, or equivalent on equated TDA).

Interpreting changes in performance

It seems clear from the data presented here that candidates of comparable ability are being awarded higher grades each year, both at A level where the trend has been consistent and substantial since 1988, and at GCSE where the change seems smaller and more intermittent. What is less clear, however, is precisely what this means.

Coe (1999), in an earlier analysis of these data, has suggested seven possible explanations for the phenomenon:

- **Teaching and learning have improved.** This is argued every year by teachers and their associations, and, indeed, may well be true. It is, however, hard to see how the claim could be convincingly substantiated. OFSTED purports to evaluate the quality of teaching and learning, but its judgements have little scientific credibility.
- **Examination performance has improved.** It could be that modern phenomena such as the increasing pressures of league tables and the frantic setting of targets might lead to an increased focus on aspects of examination preparation, without necessarily having any effect on learning. Tactics such as paying closer attention to syllabus content, endlessly practising previous years' papers or being more selective about who is actually allowed to sit the exams could all have this effect.
- **Changes in assessment have made the same levels of competence easier to demonstrate.** Changes such as the introduction of assessed coursework or the use of modular examinations could arguably have had this effect. The quality of work presented for examination may well be equal to or better than that of candidates in previous years, and therefore standards could not really be said to have fallen. However, given identical conditions, today's candidates might nevertheless be unable to match the performance of their predecessors.
- **Demographic changes have facilitated increased academic achievement.** Because of the increasing proportion of the population that can be classified as belonging to the higher socioeconomic groups, and the well know correlation between socioeconomic status and academic achievement, it is sometimes argued that this can explain the year on year rises in grades awarded.
- **The reference tests used in this analysis are inappropriate.** If performance in the ITDA or YELLIS test is as irrelevant to examination performance as shoe size, then it makes no difference how the comparison has changed.
- **The content and style of GCE and GCSE examinations have changed too much to make valid comparison possible.** Making a judgement about changes in grade standards over time requires that examinations at different times are measuring broadly the same thing. It may well be true that today's candidates would perform very poorly on an examination of ten years ago. However, if it is also true that candidates prepared for the older exam would struggle with today's papers then it is hard to argue that either is really harder; they are simply different.
- **Grade standards have slipped.**

However, even the last of these is itself problematic. Much of the annual debate focuses on whether standards have fallen, but without any great clarity about what is meant by 'standards'. Coe (2007a, 2007b) has identified three different meanings of the concept of 'comparability' each of which implies quite a different understanding of ideas such as 'standards' or the difficulty of an examination. The three are *performance comparability*, *statistical comparability* and *construct comparability*. A brief explanation of each follows.

Performance comparability

A *performance* conception of comparability locates the basis for comparing two examinations in observed phenomena such as examination scripts or coursework.

Performances in two examinations are equivalent if they are judged to exemplify the same phenomenon (or set of phenomena). Criterion referencing, in which examination outputs are judged against explicit criteria, is an example of this approach, as is the idea of comparability being based on judgements of the value of assessed outputs. In theory, *performance comparability* can be judged without any knowledge of how many people have met a particular standard.

According to a *performance* view of comparability, the 'standard' of a particular award resides in the levels of skill, knowledge, understanding – or any other qualities – that are required to achieve it. One examination would be seen as more 'difficult' than another if it required skills, knowledge or understanding that were more advanced, in other words if it made a greater demand on the candidate.

This conception often appears to be the default in thinking about comparability, though it is not often explicitly stated. When writers do not explicitly attempt to define concepts such as 'difficulty' or 'standards' it often seems to be implicit that they are thinking in terms of the intellectual demands made by an examination, the skills, knowledge and understanding that must be demonstrated to gain the award of a particular level.

Statistical comparability

The second type, *statistical comparability*, holds that two examinations may be seen as comparable if a 'typical' candidate has an equal chance of achieving a particular level in each. By contrast to the *performance* view, *statistical comparability* can only be estimated from the population of candidates; it depends entirely on how many people have achieved a particular level. However, it can, again at least in theory, be judged without the need to observe any examination outputs directly. Under a *statistical* conception of comparability, the 'standard' depends on its likelihood of being reached, possibly after taking into account other factors. An examination level is 'harder' if it is rarer, or at least estimated to be less likely to be achieved by a 'similar' candidate. Different operationalisations of this general approach include simple norm (cohort) referencing, the use of value-added models (multilevel or otherwise) and common examinee methods. However, we must be clear that the method itself does not necessarily imply a particular view of comparability; it depends how the results are interpreted.

Construct comparability

The third type, *construct* comparability, holds that two examinations may be compared if they have some construct in common. Newton (2005) discusses the extent to which one can interpret scores from two or more tests as comparable in terms of a 'linking construct'. If a plausible linking construct can be identified, it may be possible to link scores, but 'inferences from linked scores can only be drawn in terms of the linking construct' (p111). If there is a shared construct, and the same award in each corresponds to the same level of it, then they are comparable.

Construct comparability is generally demonstrated by a combination of judgement applied to observed examination outputs and statistical modelling, as, for example in equating two parallel forms of an examination, or applying a latent trait model to the results of different examinations. For this version of comparability, the 'standard' of a particular examination performance depends on the level of the linking

construct that it signifies. One examination is 'harder' than another if it indicates a higher level of the linking construct.

Conclusions

The reason any of this matters to the problem of interpreting grade drift is that whether or not standards have fallen depends on what you mean by 'standards'. If you believe that the standard is a feature of the observable examination performance (*performance comparability*), then you will trust the kind of review of standards that QCA now regularly conduct[3] or the reports of independent committees that have adopted this perspective (eg Baker et al, 2002; McGaw et al, 2004). From this perspective, it is either not really possible to say whether standards have been maintained, or, so far as one can say, the evidence suggests that they have been.

On the other hand, if you believe that a constant 'standard' means that comparable candidates should have an equal chance of achieving it (*statistical comparability*), then you may be more persuaded by the kinds of statistical analysis presented above. Quite how you interpret the statistical data is problematic, however, since the alternative possible explanations are hard to rule out on the evidence available. Students may be achieving more, but perhaps they are being better taught or prepared for examinations.

Alternatively, you may adopt a *construct comparability* perspective and interpret examination results as an indicator of some linked construct such as general academic ability. In this case, if you have a common test that measures such general ability, interpretation is more straightforward: A level grades achieved in 2006 certainly do correspond to a lower level of general academic ability than the same grades would have done in previous years. Whether or not they are better taught makes no difference to this interpretation; the same grade corresponds to a lower level of general ability.

[3] See the QCA 'Reviews of Standards over Time' at http://www.qca.org.uk/12086_1509.html

References

Baker, E., McGaw, B. & Sutherland, S. (2002), *Maintaining GCE A level standards*. London: Qualifications and Curriculum Authority. (Available at http://www.internationalpanel.org.uk/ [accessed 20.4.07])

Coe, R (1999) *Changes in examination grades over time: Is the same worth less?*. Paper presented at the British Educational Research Association annual conference, Brighton, September 1999. [Available from http://www.leeds.ac.uk/educol/]

Coe, R. (2007a) 'Commentary on Chapter 4: "Alternative conceptions of comparability" by Jo-Anne Baird' in P. Newton, J. Baird, H. Goldstein, H. Patrick and P. Tymms (Eds) *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Coe, R. (2007b) 'Common Examinee Methods' in P. Newton, J. Baird, H. Goldstein, H. Patrick and P. Tymms (Eds) *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

McGaw, B. Gipps, C. and Godber, R. (2004) *Examination Standards: Report of the independent committee to QCA*. Qualifications and Curriculum Authority: London.

Newton P.E. (2005) 'Examination standards and the limits of linking'. *Assessment in Education*, 12, 2, 105-123.

Telhaj, S., Hutton, D., Davies, P., Adnett, N., and Coe, R. (2004) 'Competition Within Schools: Representativeness of Yellis Sample Schools in a Study of Subject Enrollment of 14-16 Year Olds'. Institute for Education Policy Research, Staffordshire University, Working paper 2004/11 http://www.staffs.ac.uk/schools/business/iepr/docs/Working-paper11.doc