Why have A-level outcomes risen?

Summary

- The proportion of students receiving the highest grades at A-level has risen over the past two decades (from 11% achieving grade A in 1990, to 27% receiving A or A* in 2010). Some have suggested the modular structure of A-levels is the reason for this rise.
- A-levels are currently modular in structure. The greatest increases in outcomes occurred during the 1990s, when most A-levels were linear; so although modular assessment is not as straightforward as linear assessment in terms of grading and maintaining standards, it is not necessarily the cause of increasing grade outcomes.
- Some increases may be a result of efforts to maintain standards. The examiners' job of setting grade boundaries is a complex one, and when choosing between two adjacent marks for a grade boundary, they tend to choose the lower figure. It is a rational decision, with the knock-on effect that candidates are given the benefit of the doubt in grading decisions.
- However, as the statistical models we currently use to maintain standards rely on linking back to the previous year as the starting point for grade boundary discussions, these small 'benefit of the doubt' decisions gradually build up over time.
- In 2010 a new statistical model - 'comparable outcomes' - was introduced for A-levels which links standards back to an anchor year, rather than to the previous year. Outcomes therefore remain stable, but the model is not able to accommodate genuine improvements in performance.
- Lessons could be learnt from the Netherlands, where an established statistical model is used to inform grade boundary positioning, while maintaining standards and allowing for genuine improvements in performance.

Introduction

Concerns have been raised about the gradual improvement in A-level results and the suggestion that these rises undermine public confidence in A-level standards. It is true that the proportion of students who pass and who achieve the highest grades has risen over the past two decades (Figure 1). Understanding the causes of these increases is vital if qualification reform is to achieve its intended purpose.

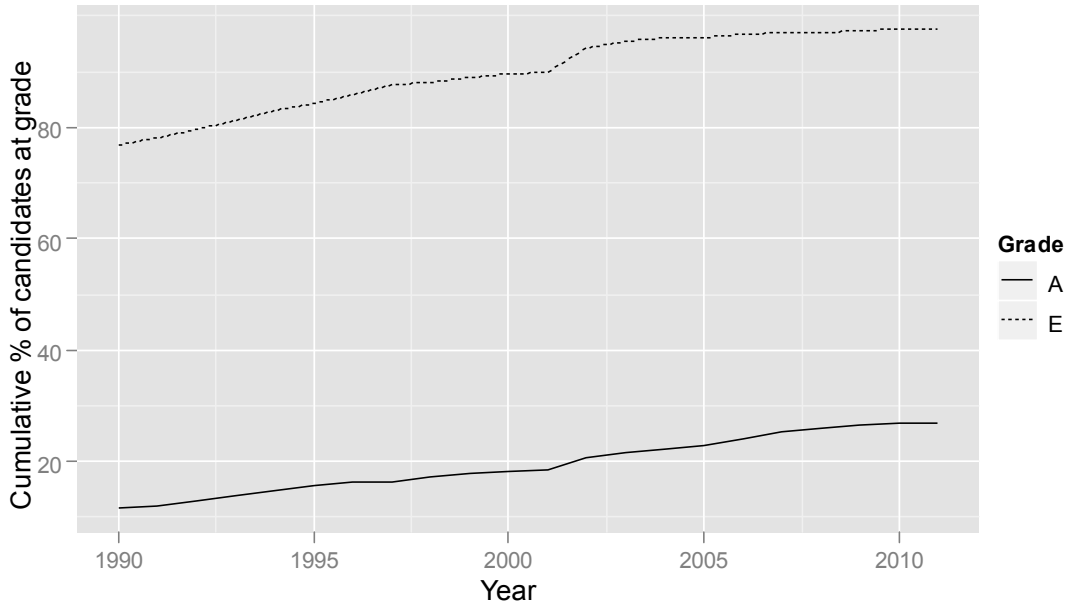

Figure 1. The percentage of A-level candidates achieving grade A and above, and grade E and above, between 1990 and 2011.

Is modularisation to blame?

Recently, it has been suggested that part of the cause of rising outcomes is the modular nature of A-levels. Indeed, modular assessment poses greater technical challenges to grading and maintaining standards than the linear equivalent, particularly at transitions between old and new specifications. These challenges can be met, as the smooth transition between specifications in 2010 shows. Also, once a new specification is established, the relationship between units and overall subject grades becomes predictable – even the effect of candidates resitting units can be accounted for – so a modular structure does not necessarily cause outcomes to increase. Moreover, the greatest increases in outcomes occurred during the 1990s, when most A-levels were linear in nature.

Nonetheless, outcomes have increased so we must ask why.

A number of hypotheses have been generated to explain the increases. They might, for example, reflect genuine improvements in teaching and learning, giving rise to real underlying improvements in student performance. At least some of the increases would most plausibly be accounted for by the emphasis on, and resources given to, raising attainment in schools. Some of the increases might, however, be attributed to the way in which we have sought to maintain standards. The Ofqual code of practice for A-level requires that grade boundaries be set on the basis of the professional judgement of senior examiners. These judgements are based on two sources of evidence: qualitative evidence, based on the careful scrutiny of candidates' work, and statistical evidence.

Each year, examiners scrutinise candidates' completed examination papers, comparing them with candidates' papers from the previous year. They are provided with written descriptions of performance at different grades and exemplars of that performance. The examiners try to take account of the relative difficulty of the two examination papers to identify grade boundary marks for the current year which are equivalent, in terms of candidates' performances, to those of the previous year.

Through research, we have learned that examiners' judgements are not as precise as has often been assumed and that, given a choice between two adjacent and similarly plausible grade

boundary marks, examiners are inclined to give the benefit of the doubt in the candidates' favour. This is perfectly rational because, as far as the examiners are concerned, the neighbouring marks each appear grade-worthy; therefore, to choose the higher of the two marks for the grade boundary would be unfair to those candidates who obtained the lower mark. Over time, these small 'benefit of the doubt' decisions can build up.

The statistical models which we used in setting grade boundaries were designed to ensure that we carried forward the standard of the examination from the previous year. In linking back to the previous year, clearly any tiny change in standards arising from the qualitative professional judgement of the senior examiners would be reflected in the statistics.

A refined statistical model, intended to overcome this issue, was introduced in the awarding of A-levels in 2010. This model is called the "comparable outcomes" approach and is now based on linking the standards back to an anchor year, rather than to the previous year. In this way, there is no risk of tiny fluctuations in standards aggregating over time and so outcomes remain stable.

The ideal solution to rising outcomes

The comparable outcomes approach clearly addresses the issue of rising outcomes, but is it the ideal solution to the question of how we set grade boundaries? Well, if we believe that the standards achieved by students – their performances in exams – do not rise over time, then it is. Likewise if the primary purpose of A-levels is still to differentiate between university applicants. This approach has the strength of containing rises in outcomes, but at the risk of not recognising real improvements in performance.

So is there a way of combining the strength of the comparable outcomes approach in supporting public confidence with the importance of recognising where we do (and equally where we do not) see real improvements in performance? There might be lessons to learn from our research colleagues in other countries.

The well-established method used in the Netherlands offers insight into how we might do this. The model used in the Netherlands allows them to establish the difficulty of an examination through the performance of candidates on hidden anchor questions that don't vary from year to year or between cohorts. Such a model offers robust information about precisely how well candidates are performing on the various aspects of that examination. If we were to introduce such a model we would be able to identify where we are seeing 'real' improvements in performance and where we need to focus resource on supporting improvements. This model would also let us compare how our students are performing with those in other countries, again identifying precisely where we are doing well and where we could improve.

If we want to see stable outcomes for our A-level examinations, we do not necessarily need to reform their structure to do this. If we want to preserve modularity or resits, for example, for educational reasons, then we now have an established statistical model which enables us to do this without further grade inflation. More refined statistical models could, in future, provide us with reassurance that outcomes were based on tangible levels of performance.

Bibliography

Baird, J.-A., Cresswell, M.J. & Newton, P.E. (2000). Would the real gold standard please step forward? *Research Papers in Education* 15, no. 2: 213–29.

Jones, B. E. (1997). Comparing Examination Standards: is a purely statistical approach adequate? *Assessment in Education: Principles, Policy & Practice*, 4:2, 249-264.

Newton, P. E. (2010). Contrasting conceptions of comparability, *Research Papers in Education*, 25:3, 285-292.

Stringer, N. S. (2011). Setting and maintaining GCSE and GCE grading standards: the case for contextualised cohort-referencing, *Research Papers in Education*, DOI:10.1080/02671522.2011.580364.

Opposs, D. (2011). *Preparation for Summer 2011 A level and GCSE Awarding*, Coventry: Ofqual. Available from: http://www.ofqual.gov.uk/downloads/category/115-board-meetings?download=1020%3A8th-june-2011-paper-24-ofqual-board . Accessed 1 May 2012.

Wheadon, C (2009). *It's a long, long time from November to June - An investigation into increasing the flexibility of delivery of high-stakes general qualifications in England through an Item Response Theory test-equating approach*, Manchester: AQA Centre for Education Research and Policy.