

'Standards are up this year' – what does this mean? The question of standards in public examinations.

Each summer, the public debate seems to commit the same logical mistake. If the numbers gaining the highest grades *increase*...then standards must be slipping (the exams, apparently, are getting easier). If the numbers attaining the highest grades *decrease*...then standards must be slipping (the attainment of students, apparently, is reducing). This kind of contradictory thinking is extremely unhelpful to all. While a degree of greater sophistication has entered the media discussion in recent years, the policy arena remains hampered by simplistic critiques of national standards within national qualifications and national testing.

A need for transparency?

The Cambridge philosopher Professor Onora O'Neill has pointed to the changing nature of public trust in major institutions and systems such as education. Her analysis suggests that no longer can a hermetically-sealed approach be taken to qualifications in general and standards in particular. Although, during the 1970s, the Schools Council embarked on a programme of improving public understanding of assessment and standards, this petered out. It was replaced by an approach amongst quangos and awarding bodies to work to maintain public confidence without embarking on any programme of open discussion of the detailed processes for running qualifications and for setting and maintaining standards. The so-called 'A level crisis' of 2002 demonstrated the inherent weaknesses of this position; whilst it was indeed a crisis of confidence, the underlying scale of the problem was far smaller than the public noise suggested.

The pressures for transparency have intensified significantly, as qualifications have increasingly been used for accountability measures. Those who have worked on issues of transparency and public understanding have explored whether we should work towards fully initiating users and the wider public into the full detail of how public assessment is run; whether we should try to enable all to understand the full technical detail; and in particular how standards are set and maintained.

A recent Cambridge seminar explored the question of just how much information is required to maintain public confidence in these processes. An analogy emerged: when people fly in an aircraft, they do not expect to have perfect knowledge of each and every system and process in and around the plane. Rather, they expect that the airline and support organisations will have complied with all necessary regulations, and that the construction and maintenance of the aircraft will have also complied with strict standards. Designing, administering and evaluating public examinations is, like aircraft engineering and maintenance, a highly technical area. So how far should we go in explaining the detail of what is done? And how far should we make clear that we are complying with strict standards, with the detail approved by a national regulator?

Defining the terms

The gross oversimplification in much press comment suggests there are some important areas which could benefit from public discussion. Perhaps the most important of these is 'what *do we mean* by 'standards'?. This isn't some esoteric exercise in splitting hairs. The logical error I referred to in the opening paragraph is all too real, and we cannot let debate continue to manifest such misleading confusions. So we need to be clear about the differences between standards of demand, standards of attainment, and content standards.

Standards of demand – the things which an assessment requires of the people who take it. When people speak of an exam 'getting easier' (or harder), this is usually the underpinning idea. There are important studies of what makes a specific examination difficult for a particular group of learners; and on things which make an examination difficult because of features unrelated to that which we really want to assess – such as confusing layout of questions, misleading examples, badly printed graphics and so on. But the key idea here – standards of demand – is of the level and nature of the skills, knowledge and understanding which are required to successfully complete the assessment.

Content standards – not quite the same as standards of demand. Content standards are associated with the value or relevance of the things which the examination includes, or the 'domains' which it is assessing. Content standards decline when an examination becomes old-fashioned, redundant or irrelevant. The content of knowledge changes over time in different subjects; the requirements of Higher Education and the economy change; the needs of society change. Qualifications thus need to change over time, in order to maintain their utility and value.

Standards of attainment – the outcomes (results) which students attain when they take an examination. If successive assessments are all at the same level of demand and the students know more or are better prepared, for example, then they attain more. Thus, the standards of attainment – overall results – improve.

I should also mention **teaching standards** and **standards of education**, since they are also confused with the forms of 'standard' above. Despite their enormous importance, this paper does not deal with them, since it is focused on assessment issues.

Standards over time

Returning to assessment, there are two more key ideas which need to be clarified: setting standards and maintaining standards. Standards need to be set when a new qualification is introduced or an existing qualification changes radically. Standards maintenance is necessary in order to make the same qualification – taken in different years by different people – of comparable standard. Different techniques are needed to support these different processes.

Some assessment specialists argue that it is unnecessary to maintain standards over time, and that it is a mistake to automatically assume this is a prime consideration.

Certainly, in some tests and examinations, the key interest may simply be who is in the 'top 25%' for any given assessment session – and the comparison between different exams and different groups in different years or on different occasions is not vital. But currently in public examinations in England, there are crucial pressures which make precise and careful maintenance of standards important in GCSEs and A levels. The first of these is national school accountability measures. The policy of judging the trajectory of improvement or decline in a school's performance according to the proportions of pupils in that school gaining specific grades in specific subjects requires qualifications to be of equal demand over time.

The second of the pressures relates to entry to Higher Education. Students who apply to university during their gap year – that is, who already have their A level results – may find themselves competing for places with those still in their last year of school, working towards their final grades. A top grade issued in 2009, then, must share a core meaning with a top grade issued in 2010 – i.e. standards of demand should be maintained and content standards held stable.

This immediately raises two key questions (other than 'how can you best do all this?'): what about standards between subjects, and what happens when qualifications change? I'll deal with these and then move on to the 'how' – the mechanisms for doing this in practice.

Standards between subjects

In the 1960s, the then Department for Education and Science looked at referencing all qualifications to a general intelligence measure, and using this to ensure that, for instance, a B grade in all public examinations 'meant the same thing'. It didn't stick – the assumptions behind it are undermined particularly by the fact that pupils are taught more effectively by some teachers in some subjects than others, and that certain subjects may particularly motivate (or de-motivate) certain types of learners. We can now add to this the fact that there are some subjects which people may encounter for the first time at A level (economics, psychology, etc), which thus contrast with subjects they have taken from a very early age (e.g. maths and English).

But accountability measures add their own peculiar pressures for 'all subjects being of the same level of difficulty'. The existence of 'hard' and 'less hard' examinations really matters if accountability measures simply look at the proportion of students gaining certain grades in different subjects – comparing, say, Media Studies with History. Is, therefore, the teacher with 30% of his Media Studies class gaining an A grade judged to be better than the History teacher with 20% A grades? It is interesting that in Australia some states simply do not worry that their post-18 qualifications are of different levels of difficulty. They take a rough gauge of how different they are (by linking the grades in all subjects back to some reference tests which pupils take earlier in their school life, and prior attainment scores) and then take this into account when looking at students who are applying to university, and come with different bundles of subjects.

In England, we are in an odd position. We know that different subjects have fundamentally different demand (work at Durham is strongly indicative of this),

notwithstanding that a mathematician finds maths easy and a geographer finds geography painless, but not necessarily vice-versa. But we treat them as being the same in crucial systems such as university tariffs and accountability measures. So what happens when qualifications change? Either the content is dramatically overhauled or the structure of the examinations changes (e.g. becoming modular/unit-based). This presents serious challenges. It would be simple if a sufficient number of people took both qualifications and had been equally adequately prepared for both – but this is not a real-life scenario. Other forms of statistical equating must be employed, and have all kinds of in-built assumptions. Equating one version of a qualification to another is not an easy matter.

Qualification change

Things become far more complex when the fundamental structure of qualifications change – such as moving from a ‘linear’ form (all examination papers taken at the end of the learning programme) to a modular or unit-based structure (with examinations taken at various times throughout the course). Awarding bodies are faced with difficult standards-setting in the first modules/units, and need to equate the final grade – after adding up all the different units, taken over time – to the previous ‘linear’ qualification. Further challenges include the very process of growing up and maturing: students may be answering questions at the end of one year where they were previously answering similar questions after two years of study. It’s not insurmountable, but it is challenging, and involves some highly technical approaches.

Just to complicate things further (as if that were necessary!), it’s not just the content of qualifications which has changed. Their very purpose has changed in the last fifty years. New purposes have come along, performance tables being one of the most notable. But look at A level: introduced in 1952 as a selection tool for Higher Education, it has remained crucial for university entrance whilst being taken by more and more pupils as a ‘general education entitlement’ – a means of giving high quality education to a larger and larger number of people, rather than simply operating as a ‘top slicing’ mechanism to find those of highest ability and attainment each year. Taken by only 6.8% of the population of 18-year-olds in 1951, it is now taken by 46.3%. Originally firmly norm-referenced, the public discussion of A level during the 1970s showed increasing concerns about the fact that the examination automatically failed 30% who took it, irrespective of whether they had enhanced attainment over the lowest 30% in the preceding years. But public memory is short and, in an age characterised by many anxieties (both genuine and manufactured), this forty-year-old moral outrage at automatic failure has, unsurprisingly, been forgotten.

Grade inflation

With the access agenda firmly established, with A level participation increasing massively, and grade attainment creeping upwards (7% gaining three A grades at A level in the mid-90s, and 17% last year), public anxiety has switched to concerns of ‘grade inflation’ and ‘debasement of currency’. Increasing access, updating content, switching to modular/unit provision – and being as transparent as possible over mark schemes, grade criteria and guidance – have all been fervent pre-occupations of policy

makers and the education establishment. Awarding bodies have delivered on that agenda. With the powerful emphasis on these imperatives, each year awarding bodies are faced with decisions about where to place the boundaries for each grade in GCSEs and GCEs. To get the right boundaries it is usually a question, for each grade, of deciding on a point in the mark scale and going up or down by one or two marks, in order to equate well with the previous years' standards of demand.

Giving the benefit of the doubt to pupils – consistent with the general moral sense of 'access' and 'best chance' which was foremost in the political agenda – can result in subtle grade inflation. Constantly enhancing the 'accessibility' of questions, the transparency of mark schemes and the precision of guidance can ease up the numbers gaining the highest grades. Changing the content to be more accessible to a wider audience than the previous educational elite can in turn move the content standards away from the precise requirements of elite Higher Education.

To investigate possible grade inflation, to publicly acknowledge that there may have been subtle drift and that a re-orientation of standards might be required, sounds like a 'Ratner moment' for awarding bodies. But it would be profoundly dysfunctional for awarding bodies to be discouraged from looking precisely and critically at the techniques and approaches they have been, and are, using. It is precisely this kind of self-criticism which has enhanced airline safety so significantly and tangibly.

Word from the top

Conversely, it is also important for awarding bodies not to be pushed into a position of blindly following orders – either out of a sense of weary resignation or as a result of excessive pressure. The orders have certainly come thick and fast: specific instructions from Government to change the content of certain qualifications, for example, or wholesale change in the shape of A levels, such as full-scale modularisation in 2000. There is a strong case for arguing that these changes have been far in excess of that which is required to keep GCSE and A level up to date, and in tune with the requirements of the labour market, society and Higher Education. And maintaining standards in times of change is technically demanding and expensive, and increases scope for subtle, unintended drift. For example, the widespread adoption and then wholesale removal of coursework (which is an important means of assessing those things produced in the course of a person's learning programme or work) have placed considerable strain on processes for maintaining standards.

The rate and nature of change in examinations –stimulated most frequently by Government action – is a crucial issue for any discussion of standards. In the face of relentless and (in some cases) unnecessary change, the processes for setting and maintaining standards have become far more complex, far more dependent on data – the principal sources of which are prior scores, subject pairs, reference tests and 'benchmark centres'. This data-dependence has effected a major, unacknowledged shift in the philosophical basis of public examinations. By using pupils' performance in earlier assessments (particularly national tests) as a reference point in setting and maintaining standards, awarding bodies have placed progressively less emphasis on what a candidate actually produces in an examination – and how that relates to things produced

by candidates in previous years. This marks a fundamental shift – a process being encouraged by the exams regulator, Ofqual (which comes into being in April 2010, but is already highly active in exams regulation). It is a subject worthy of far more debate, and is a shift which should not happen by stealth.

In fact, the situation is neither ‘Ratner’ nor ‘following orders’ – awarding bodies have responded to an extraordinarily complex set of competing demands, whilst managing the logistics of millions of scripts and results. Should the culture of press scrutiny, of public discussion and public pressure, now allow a self-critical stance – which includes public discussion of possible shortcomings of past and present examinations practice and strategy?

Science imitating art

Controversially, perhaps, I believe that assessment should be seen as an exact science. There are those who have stated that it is ‘as much an art as a science’, but I think that this is highly misleading. Some – indeed most – forms of assessment involve professional judgement, and quite rightly. Even machine-marked multiple choice have professional judgement embedded deep in them – the questions and answers have to be designed appropriately, by expert humans. But it is not true simply to say, “Because assessment involves judgement, it’s an art”. On the contrary, we need to understand with precision how any assessment is behaving; whether it manifests bias or how reliable it is. A precise and scientific ‘technology’ of evaluation and control is required to get a good grip on the measurement characteristics of any assessment, including highly judgement-based assessment of essays, of performance and so on. It may involve judgement, but we need to analyse that judgement scientifically, otherwise we have only an impoverished idea of how fair and how useful an assessment is, and indeed whether it has educational merit. This is even more important because manageable, optimal assessment often involves complex trade-offs and compromises.

Assessment must be fit for purpose, and this sometimes involves a trade-off between validity, reliability and utility. Of course we can make any assessment highly reliable – but at very high cost and administrative load. Tough decisions – underpinned by theory, evidence and consultation – need to be made in order to make an assessment truly fit for purpose. To make it something which has specific educational merit may mean some trade-offs in terms of measurement precision. Teacher assessment is, in specific settings, an approach which is fit for purpose. It may have risks in terms of precision of measurement, but is sometimes the most valid way to assess, and the approach which supports learning most appropriately.

Assessment development is not science as people crudely think of science – but then ‘pure science’ is simply not what most people think it is. Ask a research scientist: they will say that it’s messy, full of trade-offs, often compromised by the way the real world works, and yet is driven by sound method and a commitment to accuracy and evidence. Assessment is – or should be – the same.

Some questions

Where does all this leave us? Should awarding bodies operate in a context which allows self-critical review, analogous to the open ‘safety’ culture of such high-risk industries as aviation? Problems would be acknowledged, and effective remedies deployed. For example, to solve the tensions of access for the many versus the requirements of those wishing to continue to elite university study (vital for the intellectual capital of the world and the economic health of this nation), we may need more qualification routes, not fewer. We may need to avoid crude processes of fervently aligning standards in different routes where this does damage to the motivation of learners and/or the suitability of each qualification in respect of its place in the system.

The questions remain, however. Do we need to enhance public understanding of certain fundamentals? Should they be aware of the different types of standards? Should we be content to assume that ‘perfect knowledge’ of assessment is not required of them? Would it not be better for public confidence if the terms for debate were clearly defined and precisely focused? Should we not trust awarding bodies to do their job: the technical management of the development, administration and evaluation of public examinations and other assessments? Should we make it plain that, to do this, unnecessary changes to qualifications need to reduce in frequency and scope? Should standards rest on links between schools, Higher Education, employers and awarding bodies, rather than mediated through Government and its agents? Should fitness for purpose, clarity of purpose and validity be pre-eminent concerns – as emphasised in the Cambridge Approach? Without coherent answers to these questions, a more fruitful discussion of standards becomes far less likely.

Tim Oates
Cambridge
December 2009