Principles of standard setting

Lesley Meyer

1. INTRODUCTION

All awarding organisations offering general, curriculum embedded qualifications carry out standard setting[1]. Put succinctly, this is the process of establishing one or more grade boundaries for an examination, which divide the distribution of candidates' test performances into two or more categories: Pass/Fail; grade A through to grade E; grade A* through to grade G etc., depending on the type of specification involved. In ongoing examinations, the aim is to maintain standards between years, between awarding organisations and between subjects, generally in that order. Awarding organisations have to ensure that comparable examinations have the same standards. There is a defined *Code of Practice for GCSE, GCE and AEA*, written by the examination regulators[2] to promote quality, consistency, accuracy and fairness in the standard setting process. Nevertheless, there is debate each summer about public examination standards. Looking at the principles behind the standard setting process gives some indication as to why this is the case – things are not as straightforward as might be assumed.

2. THE MYSTERY OF THE EXAMINATION STANDARD

2.1 What does an examination standard 'look like'?

Right from the start, defining the term *standard* in the educational context is extremely difficult. Think about physical measurement for a moment. In 1889, the original prototype metal bar, which defined the exact length of a metre, was created in Sèvres, France[3]. Consequently, in everyday life people agree what constitutes a metre length. We can take an object which represents the length of a metre, can directly observe its length and can use it to measure the length of a second object simply by holding the exemplar against the second object and making a visual comparison. Unfortunately, we have no equivalent to measure educational attainment, nor can we develop one. It is not possible, for example, simply to keep somewhere a copy of a GCE grade A script which, when inspected, would enable us to see *exactly* what GCE grade A represents. The problem is not in retaining the script but in being able to see what it means. Unlike the metal bar, which we can see to be a metre long, with an examination script we have to interpret what is written on the page. This introduces subjectivity. Each reader will interpret the text slightly differently and therefore the standard script will represent something different to each person.

To complicate things, attainment in education is an intricate blend of knowledge, skills and understanding, not all of which are assessed on any one occasion, nor therefore exemplified in any one single script. Thus, if we ask a senior examiner to compare a script from this year's examination with an exemplar from the previous year, not only does the current script require interpretation, but generally speaking it will cover a different subset of the skills and knowledge being assessed. Further, the question difficulty and the demand of question papers will be different. Therefore, when the two scripts are compared, the comparison cannot be direct. The

[1] Standards in competency- and mastery-based qualifications are not covered by this paper.
[2] There are separate regulators in England (Ofqual, the Office of the Qualifications and Examinations Regulator), Wales (Department for Children, Education, Lifelong Learning and Skills, DCELLS) and Northern Ireland (Council for the Curriculum, Examinations and Assessment, CCEA), who jointly maintain the *Code of Practice*.
[3] at the first General Conference on Weights and Measures. The International Prototype Metre was established as the distance between two lines on a platinum-iridium bar, measured at the melting point of ice, which was designed to represent one ten-millionth of the distance from the Equator to the North Pole through Paris. In 1983, the metre was redefined as the distance travelled by light in free space in $^{1}/_{299,792,458}$ of a second. The original prototype metre is still kept at Sèvres, under the conditions specified in 1889.

senior examiner has to infer the standard of each script, and each inference is dependent on interpretation. Different examiners will place different values on the various aspects of the assessment and so conclude different things from reading the same student performance.

2.2 Why do examination standards so often hit the headlines?

Given the difficulty in defining standards in relation to education, it is not surprising that there is confusion about the way the term is interpreted and used. The fundamental problem stems from the need to distinguish between the standards of the assessment (i.e. the demand of the examination) and the standards of student attainment (i.e. how well candidates perform in the examination). Consider Figure 1, which shows the number of people reaching the summit of Mount Everest, expressed as a proportion of those who reached the summit or who died on the mountain[4]. It is unlikely that Mount Everest would have shrunk significantly in height over the short ten year period, so someone interpreting the graph would probably conclude that people have got better in climbing to the top.

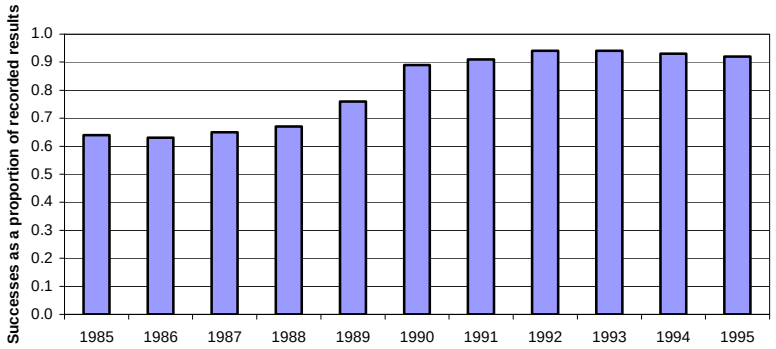

Figure 1: Mount Everest success rate variations between 1985 and 1995[5]

Now imagine that Figure 1 instead represented the national percentage of children awarded a grade C or better in GCSE[6] English over the decade. Does the graph indicate that the examination has become easier or has candidate attainment risen? With the Mount Everest data we had, quite literally, a rock to stand on; we know from experience that mountains do not suddenly change height over comparatively short time scales, so we could rule out the possibility that the standard of the assessment had been lowered, but we cannot do that in the education context. Hence the two sides in the annual argument which greets the public examination results – has examination demand reduced or has candidate attainment risen? The pass rates in themselves do not provide evidence one way or the other[7].

Consequently, whatever the summer's examination results, the press are able to generate a story about standards being in decline. If A level pass rates increase, invariably comment of

[4] i.e. the successful attempts as a proportion of the recorded outcomes. Many more people try to climb Everest each year and have to turn back, but data on the failing survivors are not readily available.

[5] Data from the Everest history website: http://www.everesthistory.com/everestsummits/summitsbyyear.htm, un-amended except for the use of a five year central rolling average (CRA) to remove annual fluctuations caused by the relatively small number of attempts made on the summit of Mount Everest in any one year. Interested readers may like to know that the successes (as a proportion of recorded results) increased again between 1995 and 2001 (the latest year for which the 5-year CRA can be calculated, given the data available at the time of writing).

[6] Or, to be pedantic, O-level or CSE, as the GCSE was first certificated in 1986.

[7] Fundamentally, examination pass rates are affected by a huge variety of factors, stemming from changes to the characteristics of the cohort year on year (for example: demographic composition; how the cohort divides between the different types of examination) or from changing curriculum characteristics (for example: examination entry policies; quality of teaching; number of teaching hours devoted to any one subject). All these factors, and more, may vary and may act to elevate or deflate pass rates even if the underlying educational standards are unchanging.

some form is made in the newspapers about the examinations being easier (see Figure 2, for example). On the other hand, if pass rates reduce, rather than being interpreted as assessment standards rising, the criticism is that the education system has failed to produce as many students who are able to climb the mountain[8].

Figure 2: ***Credit Nick Newman, Times Higher Education, The week in higher education, 27 August 2009***

Essentially, public examination results are used in a variety of different ways, many of which are inappropriate and unrealistic. It is commonly assumed, for example, that measuring the progress of educational standards over time can be achieved with reference to pass rates from examinations and that national pass rates should give us formative information about the quality of our educational system. However, pass rates tell us next to nothing about the level of attainment of candidates across the years because examination results are not designed to give us this kind of information. The demands that society makes of public examination results therefore exceed the remit of the current examining system.

2.3 How does the education profession define examination standards?

Defining the standard of a particular examination is not as straightforward as it may seem. We cannot simply identify a particular candidate's performance on a particular occasion and hold it up as the script that fully exemplifies the standard. Neither can we merely look at the pass rate. Essentially, defining the standard for an examination in a particular subject involves two things: firstly it must be established precisely *what* should be assessed by the examination; secondly, since standards represented by the same grade from examinations of the same type (GCSE, for example) should be comparable, we have to establish (at each grade) what level of attainment in this subject is comparable to that in other examinations of the same type.

For general qualifications, establishing what should be assessed is dependent on the national curriculum; as the national curriculum changes, so does what needs to be assessed by examinations. For example, over the last decade there have been continual changes to what is assessed in examinations in (what used to be called) Computing, to keep up with the continual developments in information technology. Mathematics syllabuses have been modernised. Substantial changes have been made to syllabuses in Science and Modern Foreign Languages. Essentially, what is assessed by an examination is not static and has to change to fit the needs and values of the time – not that these changes are always welcomed…

[8] Incidentally there was an infamous occasion in 1984 when the pass rates in A levels had *increased* again, causing the Daily Telegraph's headline to read "Standards falling". A week later the GCSE pass rates *fell* marginally, causing the Daily Mail to report "Standards falling"!

A-levels 'too much like sat-nav'

Professor Bailey, professor of statistics at Queen Mary, London University, told Reform researchers: "The most important change in exams over the period 1951-2008 is that sitting a mathematics A-level paper now is more like using a sat-nav system than reading a map…the result is that students will retain very little knowledge and develop very little understanding."

Katherine Sellgren (BBC News education, 17 June 2009)
http://news.bbc.co.uk/1/hi/education/8103274.stm

Requiring grades to be comparable across different subjects is another difficult issue. We want the attainment necessary to achieve a grade C in GCSE Mathematics to be comparable with that needed to achieve a grade C in English, a grade C in Economics and a grade C in Music, etc. This implies we need some way of making direct quantitative comparisons of candidates' attainments across different subjects. But this is impossible because the features which a candidate's work in different subjects has in common are insufficiently relevant to what is being assessed - we could compare the length of candidates' answers or accuracy of spelling, for example, but neither would get us very far! So to compare attainment in different subjects we are left only with *indirect* bases for comparison: professional judgement and statistics. In practice, it is these two bases upon which our examination standards are maintained from year to year.

3. MAINTAINING EXAMINATION STANDARDS IN PRACTICE

As we know, new examination papers are set in every specification each time the examination is offered and so we need to set a new pass mark (or grade boundary) for each grade. Apart from coursework[9], we cannot simply carry forward grade boundaries from one year to the next because the papers inevitably vary in difficulty from year to year and the mark scheme for one paper may have worked differently from that for the previous paper, collectively meaning that candidates may have found it easier or more difficult to score marks. To ensure the standards of attainment demanded for any particular grade are comparable between years, we have to compensate for the change in difficulty[10]. So, to determine the position of the grade boundaries, whether it be in an entirely new examination (for which the standards are being set) or an ongoing one (for which standards must be maintained), *awarding meetings* are held. In these meetings a committee of the senior examiners, who have written and overseen the marking of the examination papers - known as the *awarding committee*, or *awarders* - compare candidates' work from the current year in comparison with work archived from the previous year and in relation to any published descriptors of the required attainment at particular grades. Their qualitative judgements are combined with statistical evidence, to arrive at final recommendations for the new grade boundaries. Essentially, the role of each awarding committee is to determine, for each examination paper, the grade boundary marks that carry forward the standard of work from the previous year's examination.

The statistical evidence used in these meetings includes fairly basic data, such as the actual distribution of marks achieved and the details of the entry pattern from year to year, as well as

[9] Coursework, portfolio assessment and controlled assessments follow the same assessment criteria year on year and therefore, generally speaking, the grade boundaries are carried forward in successive years.

[10] As an aside, the annual complaints about exam results frequently surround the fact that the pass marks are adjusted each year, with claims that the grade boundaries should instead be set at fixed proportions of the mark scale (with a pass being set firmly at 50%, for example). However if an examination turned out to be far more difficult than anticipated it would be perverse to penalise students by awarding them lower grades than they would have attained had they sat the examination in any other year! Putting it very bluntly, the grade boundaries have no meaning in themselves.

more sophisticated data, often in the form of a predicted distribution of candidates' achievement. The statistical prediction tells us what we expect a cohort to achieve given their prior attainment and the relationship between the prior attainment and examination outcomes of the previous cohort of candidates (i.e. it assumes that the relationship between prior attainment and the examination outcome has remained stable between cohorts). The script scrutiny tells us whether the quality of the two cohorts' work appears comparable, i.e. whether the relationship between prior attainment and examination outcome has remained stable in practice.

GCE and GCSE awarding meetings are also informed by data comparing the outcomes of the five main awarding organisations in England, Wales and Northern Ireland offering general qualifications[11]. Each autumn, the Standards and Technical Advisory Group (STAG)[12] carry out a 'statistical screening' analysis for each GCE and GCSE specification, to investigate whether any differences exist between the outcomes of the five awarding organisations, having accounted for the ability of the candidature. If significant differences are found, a recommendation is placed on the aberrant organisation to adjust its outcomes accordingly at the appropriate grade(s) in the following summer series.

3.1 How reliable are expert judgements?

A lot of research has been carried out investigating the ability of awarding committees to carry out the task they are set in each awarding meeting, that is, to judge candidates' performances taking into account the difficulty of the examination. Unfortunately, the results indicate various problems inherent to the judgemental process.

Fundamentally, although awarders can correctly identify *whether* papers are easier or harder than those of the previous year, they are generally not good at estimating *how much* easier or harder the papers are. Awarders also find it hard to account for question paper demand, therefore candidates' performances on easy papers tend to be judged by awarders as being worthy of higher grades than those on harder papers. Even appreciating the difficulty of individual questions is not straightforward for awarders. As experts in the topic they tend to focus on the underlying *content* of a question, whereas the candidates are more likely to be distracted by the *context*. Thus, awarders may not realise that re-working a question to base it in a different framework may make it more difficult for the candidate, even if the knowledge being tested is exactly the same. Moreover, even if the awarders are made fully conscious of how easy or hard the questions have proven to be, they still cannot make a quantitative allowance for this when making their grade boundary recommendations.

There are also judgemental problems stemming from the nature of candidates' responses. Candidates tend not to perform consistently across an examination paper – they answer some questions better than others – but awarders can be influenced by the consistency of these performances, meaning that candidates who perform inconsistently across an exam paper are more likely to be considered unworthy of the particular grade than candidates who perform consistently.

The format of the current scrutiny process too creates difficulties. For example, awarders are asked to make grade boundary recommendations on each paper independently, rather than

[11] namely AQA, CCEA, Edexcel, OCR and WJEC.
[12] see the previous assessment expertise paper: 'Putting education policy into practice'

establishing the standard required to get the grade on the subject overall[13]. However, they tend to succumb to 'tunnel vision' when making judgements in this way, leading them to grade the paper more severely than they would do if they were making judgements based on students' whole performances. Also inherent to the current scrutiny process is the assumption that awarders can distinguish between candidates' work on adjacent marks. However, we know that examiner judgements of grade-worthiness are not fine-grained – they are effective in indicating whether a grade boundary mark is in the right area, but are less reliable when used to narrow the range down to a particular mark. Further, whether through vested interests or otherwise, there is a tendency for awarders to give candidates the benefit of the doubt during the boundary setting process which, unless held in check, introduces a bias towards lowering standards over time.

This isn't intended to imply that the awarders do their job badly - far from it. They are being asked to carry out an extremely difficult task and carry it out diligently year on year. The problems simply highlight that we cannot expect the awarders to do the job correctly without any other form of evidence to back up, or question, their views. That's where the statistical evidence comes in.

3.2 How reliable are the statistical data?

You'd think that an advantage of using statistical evidence would be that the data are objective, open and reliable. However, statistical data are not always as clear as we would want them to be. Just as two awarders might come up with differing recommendations for a grade boundary, so might two different statistical approaches suggest applying two different boundary marks.

There are obvious situations where statistics are more helpful than others. For example, if the entry for an examination is small - below 100, say - statistical data are very unreliable, but there is more to it than that. One of the main problems with using statistical data to define and maintain standards is that candidate attainment is affected by many factors which we cannot control, or even measure[14], so we never know, if you like, the 'complete picture'. In many cases we are able to take into account the ability of candidates taking the examination, but we are bound to assume that all the other relevant, yet uncontrollable, factors are equivalent year on year.

A specific example of this, which is an issue particularly when we are comparing standards between examinations, is that even if the overall grade distributions are identical, they may not be for particular sub-groups. It is well known, for example, that girls and boys perform differently in GCSE examinations and the relationship varies subject by subject - girls do better than boys in GCSE Chemistry but worse in GCSE Physics, for instance - so although, on the face of it, two subjects may have identical grade distributions, there may be underlying differences if we delve deeper. Whether or not you conclude that two examinations have comparable standards therefore depends on what you account for in the statistical analysis.

Assuming that historic relationships hold can be a particular pitfall of educational data. To take a basic example, we assume that centres do not change their entry policy from one year to the next, i.e. that they will stick with the awarding organisation they used the previous year (for a particular subject) and the candidates they enter will be of similar ability. Overall if the centres'

[13] There are good reasons for this, including practicability - imagine trying to get together all the work for a single candidate taking a GCE A-level or (modular) GCSE!

[14] whether related to the characteristics of the cohort (for example, candidate motivation or candidate life-styles) or to curriculum characteristics (for example, quality of teaching or number of teaching hours devoted to any one subject).

entry policy is stable, so should be the outcomes (within each centre and across all centres)[15]. However, if the entry from centres is erratic for whatever reason, expecting similar outcomes this year compared to last may not be appropriate.

Unfortunately we do not have a crystal ball to tell us what the new relationship will be. This is particularly true in the first year of a new specification, which very often has a completely different structure to its predecessor. Things are not helped by the fact that assumptions such as equivalent teaching, equivalent motivation of students, equivalent quality of textbooks etc. are far more shaky during the transition years from old to new specifications than in an established specification. Outgoing examinations can cause problems too in relation to maintaining standards, as the entry to a dying specification typically is dissimilar to earlier cohorts in various ways.

4. CONCLUDING COMMENTS: TYING THE KNOT

The setting and maintenance of examination standards is a long way from being a simple process. It is tempting to think that there should be a prototype script which encapsulates the quality of work required from a candidate to obtain a particular grade for an examination, against which we could compare a current exemplar to set the standard, but searching for it would be a fruitless exercise. The principle of carrying forward standards seems so straightforward - all we have to do is ensure that it is just as hard for candidates to get the grade this year as it was last year[16], but with so many varying factors inherent to the process, what exactly does it mean to be 'just as hard this year as last year'?

Since there is no societal agreement over what constitutes, for example, a grade A in History, arguably the examination standard can only be defined by those who have been appointed to make that decision, i.e. the awarding committee, the Accountable Officer within that awarding organisation and, ultimately, Ofqual. Moreover, since the consensus decision of an awarding committee is subjective, calling upon another group of people to ask their view of the standard that has been set is merely to add another subjective viewpoint – it would be like trying to arbitrate between fans of the Beatles and the Rolling Stones as to which was the better rock group. Essentially, while society is bound to challenge the standards of examinations (whether it is meaningful to do so or not), we are forced to accept that the standard is 'where it should be' or is 'just as hard this year as last' if the people who we trust to arbitrate the issue say it is[17].

While this may feel somewhat uncomfortable, remember that the most important requirement of the process is that candidates of similar ability in successive years receive similar grades. Candidates with similar 'potential', who are in competition for the same jobs and the same higher university places, should not be discriminated against because of the year in which they sat their exams or the awarding organisation with which they sat them. Let's be clear: in successive years, differences between specifications and in cohort characteristics are generally minimal. Stability facilitates the consistency of the awarders' annual recommendations, as does the (typically) common membership of an awarding committee from one year to the next, and also promotes statistical reliability. It is true that there is always an element of uncertainty when substantial changes occur, but that is the same in any changing scenario. Of course, there is the question of whether a better system can be devised, but until (or unless) that happens we

[15] In fact, this is quite a big assumption, as various changes may have occurred in the centre which could have affected how well prepared candidates were for the examination, but these are beyond our control.
[16] whether that relates to setting a standard for a new specification in relation to its legacy counterpart, or carrying forward standards in an established specification.
[17] just as when a church minister pronounces a couple to be married it is accepted as a social fact - no-one questions whether they really are married or not!

have to work with the existing framework. The vital thing is that we, as awarding organisations, are aware of the problematic nature of examination standards and the processes by which they are determined and do all we can to sustain quality, consistency, accuracy and fairness within the system to ensure candidates receive the grades they deserve.

5. SUGGESTIONS FOR FURTHER READING

Baird, J.-A. (2007). Alternative conceptions of comparability. *In Newton, P., Baird, J.-A., Goldstein, H., Patrick, H. & Tymms, P. (Eds.). Techniques for monitoring the comparability of examination standards (pp. 124-165). Qualifications and Curriculum Authority.*

Baird, J. Cresswell, M., Newton P. (2000). Would the *real* gold standard please step forward? *Research Papers in Education, 15(2), 213-229.*

Baird, J. & Dhillon, D. (2005). Qualitative expert judgements on examination standards: valid, but inexact. *Internal report, RPA_05_JB_RP_077.*

Baird, J. & Scharaschkin, A. (2002). Is the whole worth more than the sum of the parts? Studies of examiners' grading of individual papers and candidates' whole A-level examination performances. *British Educational Research Journal, 28(2), 143-162.*

Cresswell, M. J. (2000). The role of public examinations in defining and monitoring standards. *In H. Goldstein & A. Heath (Eds.), Educational Standards (pp. 69-104). Oxford: Oxford University Press for the British Academy.*

Eason, S. (2008). Using candidate-level GCSE results to inform on expected GCE grade distributions. *Internal report, RPA_08_SE_WP_037.*

Good, F.J. & Cresswell, M. J. (1988). Grade awarding judgements in differentiated examinations. *British Educational Research Journal, 14, 263-281.*

Impara, J.C. & Plake, B.S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions of the Angoff standard setting method. *Journal of Educational Measurement, 35, 69-81.*

Newton, P. (1997). Examining Standards over time. *Research Papers in Education 12(3), 227-248.*

Scharaschkin, A. & Baird, J. (2000). The effects of consistency of performance on A level examiners' judgement of standards. *British Educational Research Journal, 26(3), 343-357.*

Lesley Meyer
December 2009