Educational assessment standards: should they be consistent over time?

Jo-Anne Baird, Professor of Educational Assessment, University of Bristol

Uses for examination results

Whilst we can all talk about assessment standards and seemingly understand each other, it is highly likely that we will each have different ideas about standards in mind. The high jump serves as a useful analogy to describe two (of many) possible approaches. For some people, standards are related to the achievements students have to produce to be awarded a grade. In the high jump analogy, this would be the height at which the bar is set. For others, educational standards are related to the proportion of students who get over the bar. That is, the proportion of students being awarded the grade is what matters to some people.

Our notion of standards relates to the purposes to which we want to put the assessment results. If your main aim is selection for higher education, then the proportion of students being awarded the grade is very important to you. However, if you want to certify a particular level of understanding, then the requirements upon students are far more important. At the last count, Paul Newton had distinguished 22 different possible purposes for educational assessments. Each of these would have different implications for assessment design, practices and standards. All of this raises questions about expectations for assessment standards, how they are to be met and how we will know if they have been met.

Marking and setting standards are often separate processes in educational assessment at school level. With such high volumes of qualifications, it is easier to standardise the marking process and relate it to particular criteria than it is to achieve consistent grading directly. Higher education operates differently, as lecturers assign grades directly. Even where marks are used, they have a direct meaning in terms of grades that is consistent with each setting of the assessment. So, a mark of 70 or over is always considered to be a first class degree in such a system. The grade boundary for a first class degree is unchanged at 70 marks. We could discuss the extent to which this is sensible and how much information is gathered about the consistency of grading over time in higher education, but that is another topic. For school-level qualifications, we usually have a separate standard-setting process in which boundary marks are set: which students must achieve to be awarded the grades. We know that the examination papers differ in terms of how hard it is to score marks in different years, so we adjust for this by setting grade boundary marks to be higher when the examination has been easier and lower if the examination has been harder.

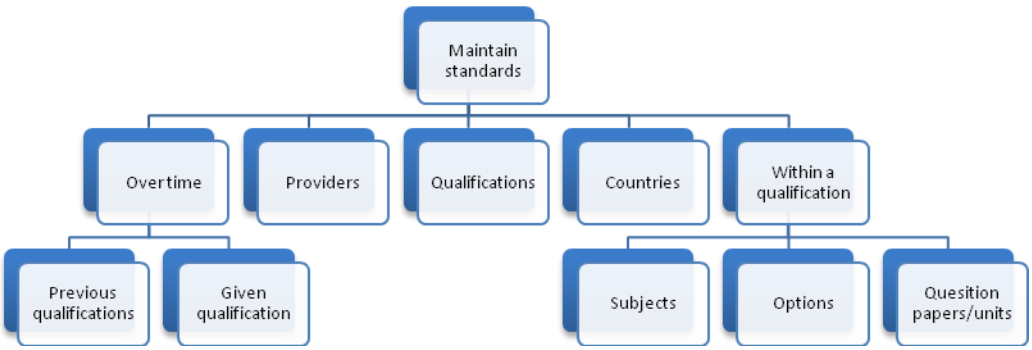
We tend to talk about the standard-setting process as being the point at which standards are set for qualifications. This is only true within certain parameters because the curriculum and question papers need to be set at the correct standard for the scale that is generated — from scoring the responses to being in the right ballpark for the right standards to be set. So, expectations for the kinds of comparability that will be delivered (and how) need to be considered in the curriculum and assessment instrument design. Having seen thousands of standard-setting meetings, there are occasional blips where the scale produced simply will not allow an appropriate standard to be set. For example, the questions in mathematics one year were so difficult that students just could not demonstrate their knowledge to a meaningful extent. We were concerned that pass marks would have had to be very low to maintain consistent pass rates with previous years which meant that what students had to do to achieve the grade would be singularly unimpressive and it would have been hard to see how it the results related to previous years' standards in terms of student performances.

Standards expectations

Often, awarding bodies have expectations for assessment standards foisted upon them from historical and cultural contexts, new policy developments and changing expectations of stakeholders. Educational assessments are not an island — they are often expected to have links with a range of other assessments. Figure 1 below might not be an exhaustive list — you might be able to think of others, but society has expectations that educational standards are equivalent in a wide variety of ways. All of this is to be achieved at the same time too!

The trouble is of course, that when you try to do this in practice and collect information about the comparability of standards, you find that there are conflicting courses of action to be taken and that all of this cannot be achieved at the same time. For example, there was a suggestion once that GCSE Chinese was too difficult when statistical techniques were used to analyse the standards. To make the standards comparable with other foreign language GCSEs, the grade boundary marks would have to be lowered dramatically. When the examination board responsible for the subject shared this with the examiners, they were horrified and explained that what remained would not constitute a GCSE in the subject because students would need to learn a tiny vocabulary and would not be able to write a sentence to pass, thereby grossly disrupting standards across years in the same subject. This example demonstrates that there are sometimes genuine educational tensions in our expectations of maintaining standards in multiple ways — in this case across subjects and across years in the same subject. Of course, this example raises another issue, which is about what we mean by standards being the same and what sources of evidence we would use. Before turning to that issue though, we need to recognise that standards over time are not the only show in town and that even if an assessment organisation successfully delivered on that expectation for comparability, there would be other societal expectations waiting in the wings from which complaints could arise.

Figure 1 Some ways in which standards are expected to be equal



Ways in which standards over time can be addressed

There are essentially two ways in which we can measure whether standards over time have changed. The first is through statistical techniques and there are a variety of those, each having different assumptions. Peter Tymms has investigated whether the standards in English schools have changed over time using a reference test technique. In this approach, the same reference test is used consistently with groups of students and we can compare the results of that test with those of the examinations that each group of students sat. Even if results on the reference test or the examinations change over time, we might expect the relationship between the two to stay constant. Let's imagine a comparison between the standard of Highers when I took them in 1986 and today's standards. To conduct such a study, a group from the 1986 cohort would need to have taken the reference test and sat their Highers and we would need to get a group of this year's students to take the reference test and collect their Higher results. If standards were consistent then, on average, people who got the same score on the reference test in 1986 and this year would have achieved the same results in their Highers. Let's imagine that people who scored 72% on the reference test in 1986 tended to get grade A in their Higher Biology examination (even though in my experience it was a tough year in 1986). We would then expect those who got 72% on the reference test this year to have gained grade A in their Higher Biology examinations, on average. Any less than a grade A on average might indicate that things had got tougher. So, statistical approaches involve holding a measure constant (in this case a reference test) and investigating whether, having controlled for that measure, the examination results have changed over time.

The other way in which we can investigate the consistency of standards is to look at the demands upon students in the examination and how well they perform. Qualitative judgements of subject matter experts are needed in this approach. So, in our example above, we would get the curricula, question papers and students' answer booklets at particular grade boundaries from the 1986 examination and from this year's examination. We would contact the subject matter experts and ask them to compare the demands upon students and their performances. Often this is a structured task in which the experts rate a variety of aspects of the demands and performances and give an overall judgement.

For completeness, I should mention that another research design involves the same students taking the two sets of tests and comparing the results. We could ask our 1986 students to take this year's examination and this year's students to take the 1986 examination in addition to their normal examination. In terms of the research design (without the benefit of time travel) we would need to ask the 40-somethings from my 1986 cohort to take this year's test. Some of us might well have forgotten a lot of our biology knowledge by now and therein lies just one problem with investigating standards over time…

Problems in measuring standards over time

So what are the others? Well, there are too many to mention more than a few here. Let's take the techniques mentioned above in reverse order and look at some of the issues.

A major problem with getting two groups of students to take two examinations is that they are unlikely to be equally well prepared for the two syllabuses and question papers. I was shocked when I sat in on an AS level standard-setting meeting only ten years after my Higher Biology examination because the syllabus was so much more modern than it had been in my day. Students were conducting DNA testing — a subject that was not even featured on TV crime investigation programmes back in 1986. Equally, today's students might not have been taught aspects of the syllabus that I followed. This makes the comparison less valid and meaningful than we would like.

Qualitative judgements of standards are notoriously difficult to make. Part of the problem is the sheer volume of material that those judging have to take into account because many qualifications involve a great deal of information about the curriculum, lengthy question papers and lots of student work. Questions about the overall standard in relation to variations on the depth and breadth of qualifications also arise. How are those who are judging to come to an overall conclusion about the comparability of qualifications if some include a wide range of material superficially and others a much narrower range of material, but in depth? We know that examiners are also influenced by the unbalanced nature of students' performances and it is the norm for students to do well on some aspects of the examination and not so well on others — often in a pattern that examiners would not have predicted. We also know that examiners are impressed by students performances on easy question papers and do not compensate their judgement enough for a bad performance on a difficult question

paper. So, asking experts to look at the quality of the work is a problematical technique, but statistics cannot wholly solve the problem either.

I mentioned above that all of the statistical techniques assume that there will be a similar relationship between a control variable (such as results on a reference test) and the examination results. However, those relationships can vary due to the selection of the groups of students that are included in the study in each year. Ideally, the groups would be representative of the cohorts who took the examinations each year, but that can be difficult to achieve. Additionally, the reference test might have more in common with one of the examinations than the other, which would again invalidate the comparison.

Theoretically and practically, there are lots of problems in maintaining standards over time. Just because our measures and techniques are imperfect does not mean, in my view, that they are worthless. We simply need to be aware of the problems and design our techniques as best we can. Given all of these problems though, you might be thinking ‘why should we bother?’ Next, I turn to an example where standards between a new and old examination were not consistent.

A salutary lesson from down under

In 2004, the New Zealand Qualifications Authority was responsible for the standard setting for new scholarship examinations. The new examinations were ‘standards-referenced’, which meant that the standards were set by expert judgement and without the use of statistical techniques. If a student performed as well as required, they were awarded the grade and vice versa, no matter what the outcomes of the examination looked like. The emphasis was upon what students needed to be able to demonstrate they could do. As the purpose of the examinations was to indicate students’ understanding and it was a new qualification, why would the examination authority need to consider consistency of standards over time? From the public inquiry that ensued, it is obvious that this was also the thinking of the examining authority officials, who were taken aback by the public outcry when the examination results were released.

Pass rates for the examination dropped by half between 2003 and 2004, with nobody passing some subjects, like physical education. In other subjects, such as Maori, the pass rate soared to be much higher than in previous years. When a major purpose of the examination results was to allocate university scholarships, this was a significant problem for stakeholders, particularly in those subjects where no one had passed. The lesson for awarding body officials is that even if you think the policies about purposes and standards are clear, you need to be aware of the societal uses to which the results are likely to be put. You need to consider whether your method of setting standards can deliver on those uses, and you need to clearly communicate the new purposes of the examinations to the public, preferably in advance of the examination results being issued.

Further reading

For a discussion of the problems with comparability in the English context:

- Newton, P.N., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (2007) *Techniques for monitoring the comparability of examination standards*. QCA publication.

Newton's 22 purposes of educational assessment can be found in Figure 1 of the House of Commons Select Committee for Children Schools and Families' Report on Testing and Assessment (2008):

- http://www.publications.parliament.uk/pa/cm200708/cmselect/cmchilsch/169/16904.htm

For an example of a review of statistical research on standards in English primary schools:

- Tymms, P. (2004) Are standards rising in English primary schools? *British Educational Research Journal*, 30, 4, 477–494.

For a description of different standard-setting methodologies with practical examples:

- Cizek, G.J. and Bunch, M.B. (2007) *Standard Setting: a guide to establishing and evaluating performance standards on tests*. Sage.