(aided by the media coverage received – partial and inaccurate as it was at times) may well have a salutary effect on the thinking of those at the highest levels involved in policy making on assessment matters. Examining Bodies have long been aware that setting standards is a difficult and complex process; partly as a result of research into comparability issues. Putting national assessments under the spotlight will make the politicians and professionals managing them acutely conscious that the concept of error of measurement has real as well as theoretical aspects.

The Project suggested some quite fundamental improvements to arrangements for national tests. Some were implemented even before formal publication of the final report and it is understood that, partly in consequence of potential risks to the maintenance of standards having been highlighted by the CoT project, a wide-ranging review of the relevant features of the key stage test system is under way. Irrespective of whether our suggestions or alternative solutions are adopted, the project has served a valuable purpose in making policy makers aware of the need to treat the conceptual and technical aspects of educational standards with greater respect.

References

Bramley, T. (2005a). 'A rank-ordering method for equating tests by expert judgement', *Journal of Applied Measurement*, **6**, 2, 202–223.

Bramley, T. (2005b). 'Accessibility, easiness and standards', *Educational Research*, **47**, 2, 251–261.

Brooks, G., Foxman, D. & Gorman, T. (1995). *Standards in literacy and numeracy,* New Series 7. London: National Commission on Education Briefing.

Christie, T. and Forrest, G. (1981). *Standards at GCE A-level: 1963 and 1973*. London: Schools Council Publications/Macmillan Education.

Cresswell, M. (2000). 'The role of public examinations in defining and monitoring standards', *Proceedings of the British Academy*, **102**, 69–120.

Goldstein, H. (1983). 'Measuring changes in education over time: problems and possibilities', *Journal of Educational Measurement*, **20**, 4, 369–377.

Massey, A. (1994). 'Standards are slippery', *British Journal of Curriculum and Assessment*, **5**, 37–38.

Massey, A. and Elliott, G. (1996). *Aspects of writing in 16+ English examinations between 1980 and 1994*, Occasional Research Paper 1. Cambridge: UCLES.

Massey, A., Green, S., Dexter, T., and Hamnett , L. (2003). *Comparability of national tests over time: key stage test standards between 1996 and 2001*, final report to the QCA of the Comparability Over Time Project. Available from http://www.qca.org.uk/news/6301.html.

Newbould, C. and Massey, A. (1979). *Comparability using a common element*, Occasional Publication 7. Cambridge: TDRU .

Newton, P. (1997). 'Examining standards over time', *Research Papers in Education,* **12**, 3, 227–248.

Statistics Commission (2005). *Measuring Standards in English Primary Schools, Report No 23*, February 2005. London: Statistics Commission.

Tymms, P. (2004). 'Are standards rising in English primary schools?', *British Educational Research Journal*, **30**, 4, 479–493.

STANDARDS OVER TIME

Accessibility, easiness and standards

Tom Bramley Principal Research Officer, Evaluation & Validation Unit

The following is a summary of a research article published in summer 2005 which was prompted by my experience of working in the National Curriculum test development group, formerly part of UCLES' Research & Evaluation Division, now part of OCR. One major task in the test development process is to carry out statistical and judgemental exercises which can provide evidence about where to set the cut-scores on the test. These cut-scores (level threshold boundaries, equivalent to grade boundaries on GCSEs and A-levels) are supposed to be at the same standard each year. Of course, tests can vary in difficulty and the cut-scores might not be at the same point on the raw mark scale each year, in order to allow for fluctuations in difficulty of the test from year to year (as with GCSEs and A-levels).

However, discussions about whether one year's test is easier or more difficult than the previous year's test can often get bogged down when the spectre of 'accessibility' raises its head. Is a 'more accessible' test the same as an 'easier' test? Are there any implications for where the cut-scores should be set if a test is deemed to be more accessible, as opposed to more easy? Is there any way to identify questions which are 'inaccessible'?

The main purpose of the article was to use a psychometric approach to attempt to answer these questions. The article begins by discussing the meaning of 'standards' and the ambiguity with which the term is used, particularly in media reporting of examination issues. The standard can be defined psychometrically as a point on the latent trait which is assumed to underlie or cause the responses to the test questions. The informal definition of statistical equating – that if standards have been correctly applied to two tests then it should be a matter of indifference to candidates whether they take test A or test B in terms of which level they obtain – is used as a starting point for discussing the issues raised by accessibility.

Three prototype arguments in favour of not raising the cut-scores by as many marks as the statistics might suggest when a test is deemed to be more 'accessible' were used to illustrate the discussion:

> *The paper is more accessible, but the amount of science hasn't changed.*
>
> *We've removed some of the hurdles which prevented the pupils from showing us what they can do.*
>
> *The pupils will be less 'turned off' by the paper and so we'd expect performance to improve.*

I have heard variations of these arguments on many occasions in the course of my time at UCLES – you might wish to pause here to consider whether you think they are valid.

In the article I argued that all of these arguments ignore the basic idea behind statistical equating that it should not matter to the pupil whether they take the more accessible or less accessible test. From a psychometric perspective, these arguments are all saying that the new accessible test is not measuring quite the same trait or construct as the old, less accessible test. To the extent that this is true, then strictly speaking it is not possible to set cut-scores on the two tests which have the same meaning.

However, it is at the level of the item/question that the actual changes in accessibility occur. The second part of the article explores the use of Rasch misfit statistics to investigate accessibility issues within a single test (i.e. it simplifies the situation from the actual one which occurs in practice where we have two different tests). It turns out that both types of misfit (underfit, and overfit) have the potential to diagnose problems with accessibility. Substantial underfit (lack of discrimination) often indicates either a poorly worded or ambiguous question which has confused the more able pupils, or an incorrect or incomplete mark scheme. Substantial overfit (discrimination which is too good!) might signify a different, but highly correlated dimension. For example, a teaching effect could produce overfit. If only the most able pupils in a school are taught a particular topic then only they will be able to answer a question on that topic correctly. This would produce good discrimination, but this good discrimination is arguably invalid, because the question would be inaccessible to the lower ability pupils.

Of course, such psychometric indicators are only the starting point for qualitative research aimed at deducing patterns and rules for identifying particular types of question and response format where accessibility issues are likely to cause a measurement problem.

I concluded the article by asserting that treating the standard setting/maintaining issue as a measurement problem provides a rational basis for understanding accessibility. Increasing accessibility does in fact make the test easier and cut-scores should rise in order to maintain standards.

Further reading

Bramley, T. (2005). 'Accessibility, easiness and standards'. *Educational Research*, **47**, 2, 251–261. Available from http://www.tandf.co.uk

STANDARDS OVER TIME

A rank-ordering method for equating tests by expert judgement

Tom Bramley Principal Research Officer, Evaluation & Validation Unit

The following is a summary of a research article published in summer 2005.

This paper built on much research carried out at UCLES over the past ten years on the use of judgements in scale construction. The main technique which we had used was Thurstone's paired comparison method whereby two objects are compared in relation to a single trait. Repeated comparisons of different pairs of objects by several (or many) judges can allow a single scale to be created, with the objects located at different points according to how many comparisons they 'won' and the location on the scale of the objects with which they were compared. Thurstone's method has been used for comparability exercises (comparing scripts at the same grade boundary from the same subject at different exam boards), and for research into the perennial question of standards over time (comparing scripts at the same grade boundary in the same subject specification in different years).

As those who have been involved in these studies can testify, the Thurstone approach can be very time-consuming and tedious for the panel of judges involved, because of the number of judgements required to form a satisfactory scale. My idea was to attempt to speed up the process by asking judges to place a set of objects into a single rank order, rather than requiring many separate paired comparisons. A second variation on the Thurstone process as it had been used in the studies above was to involve the entire mark range, rather than focussing on a particular boundary. This was to allow the two mark scales to be compared at all points by plotting the mark on the script against the 'judged measure' (the outcome of the ranking procedure).

NAA (the National Assessment Agency, responsible for the National Curriculum tests) requires its test development agencies to carry out a judgemental standard-setting exercise using practising teachers in order to supplement the statistical procedures used to derive cut-scores on the current year's test. The agencies are given some flexibility in the methods they use, so we decided to try out the rank-ordering method in this context. Scripts from the Reading component of the Key Stage 3 English test in 2003 (live test scripts) and 2004 (final pre-test scripts) were used as the objects to be ranked. Approximately 40 scripts from each year in total were involved in the exercise, one on each mark covering the effective mark range of each test. All the question mark totals were 'cleaned' from the scripts so the judgements would be based on perceived quality rather than simply adding up the marks.

Each judge (from a panel of twelve) was given four packs of ten scripts. Each pack contained five scripts from 2003 and five from 2004. No two packs of scripts were identical, but there was a lot of overlap across judges and packs in order to create the linking necessary to form a single scale from their judgments. The judgemental task was simply to put the ten scripts in each pack in order from best to worst. Tied ranks were allowed, but strongly discouraged, and in the event there were only two