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 or three tied rankings in all 48 rank orders. The contents of each pack were systematically varied in terms of both the overall level and spread of scripts from each year. The judges were warned not to make any assumptions about the contents of their packs – it was possible (for example) for all five scripts from one year to be 'better' than all five scripts from the other.

The data were analysed with two statistical methods, both based on the Rasch model. The first method converted each ranking into a set of paired comparisons and proceeded to analyse them as usual. The second method treated each ranking as a separate Rasch Partial Credit item. When the resulting measures from the two methods were plotted against each other the points lay on a straight line, showing that the two methods were giving substantively the same result.

More interesting was the outcome of the exercise, obtained by plotting the mark on the script against the judged measure, and fitting separate best fit lines for each year, as shown in Figure 1.

Since the judged measures are all on the same scale, the two raw mark scales can be equated (perhaps a weaker term such as 'linked' is more appropriate): the marks corresponding to the same measure are deemed to be equivalent. The equivalent mark on the 2004 test to any mark on the 2003 test can be found either by reading off the graph, or by using the regression equations for the best fit lines. In fact, in this case the two best fit lines were approximately parallel, separated by a vertical distance of around three marks, leading to the conclusion that the 2004 Reading component was about three marks easier at all levels than the 2003 Reading component. This agreed well with the (completely independent) evidence from statistical equating of pre-test scores, which had suggested that the 2004 test was around two marks easier.

The article contains a lengthy discussion of the difference between standard setting and standard maintaining, arguing that the rank-ordering method is more appropriate than most other judgemental methods for standard maintaining, and that standard maintaining is more appropriate than standard setting in the national testing context.

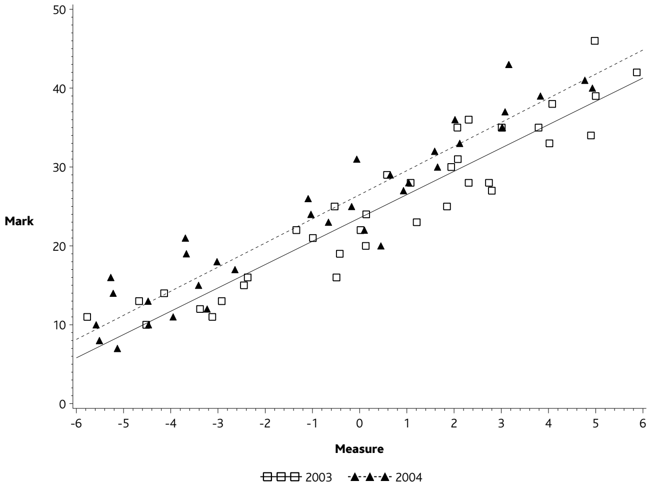

Figure 1: Plot of mark against measure for scripts from 2003 and 2004.

Since the paper was written, the method has been repeated successfully with the 2004 and 2005 Writing components of the KS3 English test, and is currently being investigated in a research study using scripts from two years of an A-level Psychology paper. There are also plans to investigate its suitability as an award meeting methodology.

Further reading

Bramley, T. (2005). 'A rank-ordering method for equating tests by expert judgement', *Journal of Applied Measurement* **6**, 2, 202–223. Available from http://www.jampress.org

DESCRIBING ACHIEVEMENT

A review of research about writing and using grade descriptors in GCSEs and A levels

Dr Jackie Greatorex Principal Research Officer, Research Programmes Unit

In this article I describe current Awarding practice and review some of the literature about writing and using grade descriptors (often also referred to as 'grade descriptions') for GCSEs and A-levels. Particular emphasis is given to the research that has used empirical evidence to write grade descriptors and the associated research methods.

> *Grade descriptors are descriptions of the qualities expected at different levels of a candidates' performance in an assessment* (Greatorex et al., 2001, 167).

The following are some extracts from grade descriptions for GCSE Biology:

Grade F: Candidates recall a limited range of information. For example, they state the main functions of organs of the human body and describe some defence mechanisms of the body (*OCR, 2000, 17*).

Grade C: Candidates describe how evidence is used to test predictions made from scientific theories, and how different people may have different views on some aspects of science (*OCR, 2000, 18*).

Grade A: Candidates use detailed scientific knowledge and understanding in a range of applications relating to scientific systems or phenomena. For example, they explain how temperature or water content is regulated in humans (*OCR, 2000, 18*).