Educational Research, Vol. 47, No. 2, June 2005, pp. 251–261

Routledge
Taylor & Francis Group

Accessibility, easiness and standards

Tom Bramley*
Assessment Directorate, UCLES, Cambridge, UK

In setting the cut-scores on National Curriculum tests it is important to maintain standards. In the process of test development, both within and across years, changes are made to the style of the questions in order to increase their 'accessibility'. This raises the question of whether a more accessible test should have higher cut-scores. Purely statistical definitions of equating are blind to differences between 'accessibility' and 'easiness' and cut-scores derived from statistical equating methods will be higher for a more accessible test. Arguments about the increased validity of the more accessible test are sometimes used to justify not raising the cut-scores as much as would be indicated by statistical methods. These arguments are shown to be equivalent to postulating that changing the accessibility is changing the construct measured by the test. Using a statistical measurement model can provide a rational basis for understanding accessibility and identifying types of question where accessibility issues are causing a measurement problem.

Keywords: *Standards; Equating; Accessibility; Measurement; National Curriculum test; Rasch*

Introduction

In National Curriculum (NC) testing in England, pupils take statutory tests at the end of each of three 'key stages', when they are aged 7, 11 and 14 years, respectively. At key stage 1 (KS1), English and mathematics are assessed. At KS2 and KS3, science is also assessed. The results are used to monitor national attainment, to evaluate the impact of political and educational initiatives, and (for KS2 and KS3) to compile 'league tables' of school performance. A completely new version of each test is produced each year, so one critical part of the test development process is to set cut-scores on the current year's test which represent the same standard as the cut-scores set on previous years' tests.

Definition of cut-score

Each pupil taking a NC test will achieve a 'level' on the NC scale within the range of levels available for that particular test. Each 'cut-score' is a point on the raw score scale of the test which separates the pupils with raw scores around that point into pairs

*Assessment Directorate, UCLES, 1 Hills Road, Cambridge CB1 2EU, UK.
Email: Bramley.T@UCLES.org.uk

ISSN 0013-1881 (print)/ISSN 1469-5847 (online)/05/020251-11

DOI: 10.1080/0013880500104382

of adjacent levels: for example, if the level 4 cut-score is at 40 marks, those with less than 40 marks will be categorized as level 3 and those with scores of greater than or equal to 40 marks will be categorized as level 4. Pairs of cut-scores therefore define ranges of marks that correspond to each available level. For example, if the level 4 cut-score is at 40, the level 5 cut-score at 76 and the level 6 cut-score at 107, then pupils with raw scores from 40 to 75 will be categorized as level 4, and those with raw scores from 76 to 106 as level 5.

Definition of standard

'Standards' is a word with several meanings that are often confused in debates about education. In the summer of 2002 there was widespread concern about the way cut-scores had been set on the A-level examinations produced by the three awarding bodies in England and Wales, prompting the UK government to commission an independent inquiry - led by Mike Tomlinson, the former head of the Office for Standards in Education (Ofsted), a UK government department responsible for the inspection of all schools in England). In his interim report (Tomlinson, 2002), he stated:

> At the root of this is a longstanding misunderstanding of the difference between maintaining a standard and the proportion of candidates meeting that standard ... This misunderstanding appears to exist at almost all levels of the system, and in society at large.

This paper is written from a psychometric viewpoint, where a test score is seen as an indicator of a pupil's position on an abstract or 'latent' trait (e.g. Lord, 1952). This trait, sometimes referred to as a variable or construct or dimension, is conceived as a continuum analogous to physical dimensions like length or temperature. The standard is the point on this latent trait that marks the boundary between two meaningfully different categories (e.g. pass/fail, A/B, level 5/level 4). This standard exists independently of the test itself and therefore has to be applied to each new test that is constructed.

This should not be confused with a common misconception of the meaning of standard, namely the proportion of some population who meet or exceed a standard. For example, if the proportion of pupils who pass a test rises from one year to the next, this can be interpreted as meaning that standards are rising, that is the general level of the trait in the population. This interpretation, of course, requires that the standard in the sense defined above has been correctly applied to both tests. If it is not believed that this standard has been applied correctly to both tests, then a rise in the proportion of pupils passing a test could also be taken as evidence that standards are falling, as is often seen in newspaper reporting of GCSE and A-level results.

To return to the task of setting cut-scores on National Curriculum tests, the aim is to ensure that the standard set in previous years is applied to the current year's test when setting the cut-score at each level.

Why do the cut-scores need to change?

If the current year's test is out of the same number of marks as the previous year's test, and has been carefully designed to meet the same specification of content and difficulty, one might assume that there is no good reason to change the cut-scores from one year to the next. In practice, however, it is almost impossible to produce two tests that are exactly the same. In particular, it is recognized that the current year's test might be more or less difficult than the previous year's test, and that the cut-scores might therefore need to be different to allow for this.

In general terms, if the current year's test is thought to be more difficult than the previous year's test (by the panel of decision-makers who decide where the cut-scores should be set), it will be thought necessary to lower the point at which the cut-scores occur on the overall mark scale. If the current year's test is thought to be less difficult than the previous year's test, it will be thought necessary to raise the point at which the cut-scores occur on the overall mark scale. Often it is thought necessary to make different adjustments at different levels.

But what exactly is 'difficulty'? It turns out that this property of a test can only be defined simultaneously with another property of the pupils who take the test, namely their 'ability'. Use of the word 'ability' is often contentious, but here it does not refer to some innate ability, aptitude, potential or IQ. It is simply used in the neutral psychometric sense to refer to a pupil's location on the latent trait.

Similarly, 'difficulty' is simply the location of a test question on the same latent trait. The relationship between ability, difficulty and observed performance is expressed mathematically in what are known as Latent Trait Measurement Models. The simplest of these, the Rasch model (Rasch, 1960), expresses the probability of success on a question as a function of the difference between pupil ability and question difficulty. The Rasch model is mathematically equivalent to the simplest Item Response Theory (IRT) model, although it is derived from a different perspective (see e.g. Andrich, 1989). More complex IRT models add extra item parameters (e.g. to represent discrimination and guessing) but all conceive of the test as locating both people and questions on a single dimension of measurement. Both ability and difficulty are estimated simultaneously from the observed performance data—the mark obtained on each question by each pupil.

So, if the test is held constant, a pupil with more ability (i.e. at a higher location on the latent trait) will be expected to gain more marks than a pupil with less ability. If the pupil is held constant, they will be expected to gain fewer marks on a more difficult test (i.e. a test containing questions with difficulty parameters which are higher on the latent trait) than on an easier test. The location of the standard on the latent trait does not change, but will correspond to different raw score totals on tests containing questions of different difficulty.

Statistical equating

One informal way of expressing the goal of setting cut-scores would be that it is trying to ensure that pupils with the same level of ability will be awarded the same National

Curriculum level on the current year's test that they would have been awarded on a previous year's test. This, of course, requires the assumption that the two tests are testing the same thing, or similar enough things for it to be meaningful to say (for example) that a level 5 on one year's test 'means the same' as a level 5 on another year's test. A different way of expressing the same goal is that it should be a matter of indifference to candidates at all levels of ability which of the two tests they take (a paraphrase of Lord's 'equity condition' for equated scores (Lord, 1980)).

There are many different possible methods for equating two tests by statistical means. Different methods may require different data collection designs and may involve different assumptions (a comprehensive reference is Petersen *et al.*, 1989). In fact some different methods are used in National Curriculum testing across subjects and key stages. However, all assume both that the task is meaningful, and that the fundamental goal is as described informally above.

Accessibility

One issue that arises in the course of test development and the setting of cut-scores is that of 'accessibility'. In the course of test development, both within a year as questions are pre-tested and refined, and in more general terms over the years as educational policies and goals evolve, there is a concern that the questions should be 'accessible' to pupils. This is a particular concern for teachers, who want the testing experience to be as meaningful as possible for all their pupils, given that the tests are compulsory and (arguably) not as 'high stakes' for the pupils as for the school. It is interesting to note straight away that the term used is 'accessible' and not 'easy'. An increase in accessibility can be achieved by means such as simplifying the language in the questions, introducing more prompts, improving clarity of layout and diagrams, and changing the balance of the different response formats to reduce the number of questions requiring one or more written sentences and to increase the number of questions requiring one-word answers or requiring the pupil to select from a given set of possible answers.

In a comprehensive investigation into the comparability of standards set on NC tests over time, Massey *et al.* (2003) considered the issue of accessibility. One strand of their research was a survey of pupils' perceptions of the test materials. Even the youngest children were able to distinguish features of the tests that had changed over time and to express their preference, with reasons, for the later versions of the tests (specifically, those of 1999, 2000 or 2001 compared with 1996). Massey *et al.* noted that while some of these changes in accessibility might increase pupils' engagement with the tasks, or reduce stress, they would not necessarily increase the probability of producing the correct answer. On the other hand, they assumed that other changes, designed to increase accessibility by making the tests easier to understand and respond to, *would* increase the probability of producing the correct answer.

The question then, of course, arises of what effect an increase in accessibility should have on the cut-scores. There is no doubt that the scores obtained on the more accessible test will be higher, for a group of pupils with the same distribution of ability. The debate is whether the cut-scores should also be higher. Massey *et al.*

(2003) note the tension between the desire to increase test validity by giving pupils the chance to show what they know and can do, and the need to maintain comparability over time in the sense of it being a matter of indifference to pupils which year's test they take in terms of what level they achieve:

> If we change the tests, so that more children are able to demonstrate that they have the qualities we seek to reward, does this not also change test standards? Could we live with a testing system where equivalent children would get different results in different years as accessibility varied?

The main purpose of this paper is to show how a psychometric approach can clarify the issues surrounding the effect on cut-scores (and standards) of changes to accessibility. The first thing to note is that distinctions between easiness and accessibility are in the mind of the humans debating the issue. The statistical methods used in equating one test to another are blind to this distinction, and so will treat a more accessible test as an easier test, and produce higher cut-scores to maintain the standard. That is, the features of the questions which lead to higher (or lower) scores on the test (e.g. intrinsic difficulty, accessibility, familiarity) are all lumped together under 'difficulty'. Changing any one of these features will alter the position of the question on the latent trait and thus affect the difficulty parameter of the question. Similarly, the features of the pupils which lead to higher (or lower) scores on the test (e.g. intrinsic ability, quality of teaching, motivation, revision) are all lumped together under 'ability'. Changing any one of these features will alter the position of the pupil on the latent trait and thus affect the ability parameter of the pupil in question.

It is in fact very difficult to explain these latent variables (e.g. intrinsic difficulty, intrinsic ability, motivation, accessibility) without circularity and yet they are so ingrained into the ways of thinking and vocabulary of assessment that practitioners are inclined to talk about them as though everyone knows what they are, and how to measure them. But it is worth noting that the theoretical status of such variables and their role in explanations of observable phenomena is currently a topic of academic debate (see e.g. Borsboom *et al.*, 2003). Unless a way is found unambiguously to determine their separate and combined effect, then many arguments that draw on these terms can easily become unsupported assertions, or even statements of faith.

Three examples of arguments that could be used to support *not* raising the cut-scores when a test is made more accessible (or, more realistically, not raising them quite as much as would be suggested by the statistical methods which are blind to the distinction between easiness and accessibility) are listed below. Science tests are used as the example, but the arguments would apply to any subject:

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

Arguments 1 and 2 are essentially saying that the test is now a more valid test of the same material. It is implicitly ignoring the idea behind statistical equating that it should not matter to the pupil which test they take and instead saying that the more accessible test is a more valid test of all the pupils' abilities, because the scientific content is the same, but some science-irrelevant hurdles have been removed. Argument 3 is different in arguing that a group of pupils with the same distribution of ability will have their ability increased by extra motivation from the accessible test. Although this argument focuses on the pupils rather than the accessibility of the questions, it also ignores the idea that it should not matter to the pupil whether they take the more accessible or less accessible test.

Can the issue be resolved?

One way to deal with the problem is to recognize that the whole issue of standards is essentially a subjective one, and that standards ultimately reside in the values of the particular community of 'users', or people who take decisions or make judgements based on the outcomes of a particular assessment. Statistical methods are not properly applicable to these values. This approach has been explained in depth by Cresswell (e.g. Cresswell, 1996). If thought appropriate in this context (National Curriculum testing), then this approach could justify not raising the cut-scores to the extent suggested by statistical methods when a test has been made more accessible.

However, it can be argued that Cresswell's approach is much more appropriate when considering a broader context such as whether an 'A' grade at A-level Physics means the same as an 'A' grade at A-level Business Studies; or whether a 'C' grade at the old O-level means the same as a 'C' grade at GCSE. This is because it is much less obvious that the tests being compared are in fact measuring the same thing in these broader contexts. In the much narrower context of comparing two tests designed to the same specification in the same subject in consecutive years, there is a stronger case (as Cresswell, 1996, acknowledges) for trying to use relatively objective arguments based on psychometric and statistical reasoning.

In the example above, the three arguments in favour of not using the statistical definition of equating as a basis for cut-score setting can all be viewed from a psychometric perspective as saying that the new accessible test is not measuring quite the same dimension or construct as the old less accessible test. For example, the old less accessible test may have been measuring a dimension of 'reading' and/or 'writing fluency' (hereafter referred to as 'literacy'), as well as science ability, and the increase in accessibility may have removed the presence of this extra dimension in the test (arguments 1 and 2). Also, the less off-putting nature of the questions may have removed a 'motivation' dimension from test performance (argument 3).

But how can these extra dimensions be detected? Do they exist purely in the minds of the beholders? In principle, it might be possible to vary experimentally the hypothesized science and literacy content of a test, and administer it in conditions of controlled levels of motivation to randomly equivalent groups of pupils. In practice, simply not enough is known about these variables and how they interact to achieve the

level of 'stimulus control' (Michell, 1990) necessary to carry out such an experiment. In practice, therefore, different dimensions tend to be detected by internal correlational techniques such as factor analysis or principal components analysis. On such an approach, two conceptually different dimensions that correlate perfectly with each other cannot be empirically distinguished. That is, if (for example) literacy is perfectly correlated with science ability, then they are in effect a single dimension, and making a test easier in terms of its literacy demands is the same as making the science easier. If the same group of pupils took both tests, the rank order would be exactly the same, but scores on the more accessible test would be higher, and the cut-scores would be correspondingly higher.

On the other hand, if literacy is completely uncorrelated with science ability, then making a test easier in terms of its literacy is effectively changing what the test is measuring. Recall that the first prerequisite for any meaningful equating is that the two tests being equated are measuring the same thing. Changing the balance of the two dimensions would make it impossible (strictly speaking) to attempt to set cut-scores on the more accessible test which had the same 'meaning' as the cut-scores on the first test.

Misfit to a measurement model

One way to tackle the problem of identifying these extra dimensions is to consider changing a particular test to make it more accessible (i.e. using the same questions, but altering their accessibility). This simplifies the situation from the context considered earlier, where one year's test contains different questions that might also be more accessible than the previous year's test. It is important to note that changes to accessibility will not apply equally to all questions—they will apply to specific questions. How can questions that might need to be made more 'accessible' be identified?

If the data from the original test are analysed using the Rasch measurement model (1-parameter IRT model), then for each question an 'item characteristic curve' (ICC) will be obtained. This shows how the probability of a correct response increases as ability increases, according to the model. If the test consists entirely of 1-mark questions (which for the sake of clarity it will be assumed it does), then the ICCs for all questions will have the same slope (i.e. will be parallel), just differing in location. Although this is sometimes regarded as a limitation of the Rasch model, in fact it is necessary for the data to conform (within stochastic limits) to this requirement in order to be able to construct unambiguous measures of person ability and item difficulty from the data (see e.g. Wright, 1999).

If the empirical ICC is plotted (e.g. by dividing the pupils into four groups according to ability and plotting the mean score of each group against the mean ability of each group) on the same chart as the modelled ICC, then questions which are not functioning in the same way as the majority of the other questions can be diagnosed. Such questions are described as 'misfitting' in the sense that responses to these questions do not match the model expectations, given the ability of the pupils and difficulty of the questions estimated from the data.

There are two types of misfit that might be described as 'underfit' and 'overfit'. For an underfitting question the empirical slope will be significantly less (shallower) than the model slope. Such a question is illustrated in Figure 1. This is a question that would be identified as having low discrimination in a traditional item analysis. An overfitting question, by contrast, has an empirical slope that is significantly greater (steeper) than the model slope. Such a question is illustrated in Figure 2. This question would have a high discrimination in a traditional item analysis and would therefore appear to be an ideal question.

Substantial underfit (lack of discrimination) often diagnoses a poorly worded or ambiguous question which has confused the more able pupils, or an incorrect or incomplete marking key, and such questions are rightly scrutinized first by test developers. This is because those identified as lower ability by the model have performed better than expected (in Figure 1 their circles are above the line) and those identified as higher ability have performed worse than expected (in Figure 1 their circles are below the line). An underfitting question might be also taken as measuring

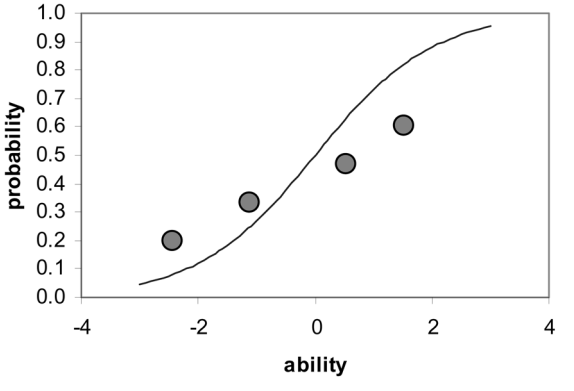

Figure 1. A low discrimination question (underfit)

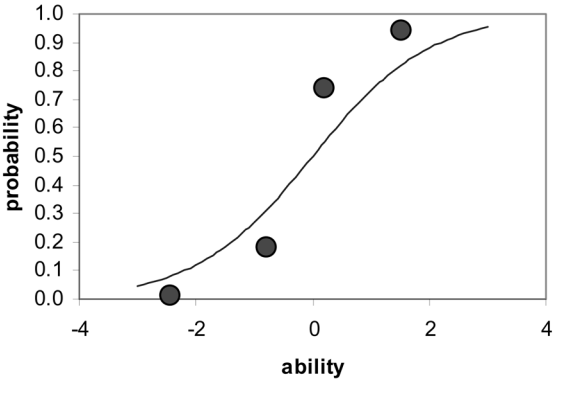

Figure 2. A high discrimination question (overfit)

a significantly different (uncorrelated) dimension to that accounted for by the model (i.e. the ability/difficulty dimension). This might be the case if (for example) a certain science question really tested everyday knowledge, and if it is true that everyday knowledge is not highly correlated with science ability.

An overfitting question, on the other hand, might signify a different, but highly correlated, dimension to that accounted for by the model (Masters, 1988). For example, if only the most able pupils are taught a particular topic, then only they will be able to get a question on that topic correct. The pupils identified as lower ability by the model will perform (even) worse than expected (in Figure 2 their circles are below the line) and those identified as higher ability will perform (even) better than expected (in Figure 2 their circles are above the line). The dimension of 'teaching' has interacted differently with ability on this question to the other questions. A second example might be if one science question on a test called for a higher level of literacy than the other questions. If literacy is highly correlated with general science ability, then this will create an overfitting question because only the more able pupils will be able to 'access' the question.

Thus overfit can also be used to diagnose questions that could need altering, despite their having a high discrimination and appearing to be functioning well in traditional item analysis terms. If a change is made to such a question which increases the chance that the lower-ability pupils will get it correct, the effect will be that the curved line in Figure 2 will shift to the left (because the question as a whole is easier), bringing the upper circles onto the line, and the lower circles will rise (e.g. because the reading and writing demands which hampered the lower-ability pupils has been reduced), bringing them on to the line too. The question will have a slightly lower discrimination, but a better fit!

This example indicates how the problem of accessibility might be approached by empirical means. It may well be that demands on literacy (or other sources of 'hurdles') can manifest themselves in both types of misfit. By analysing both the questions identified as underfitting and overfitting, it should be possible to discern patterns and deduce rules for identifying particular types of question or response format where accessibility issues are likely to cause a measurement problem. The difficulty of such a task should not be underestimated and the psychometric analysis only provides a starting-point for detailed qualitative research. The work of Pollitt and colleagues (e.g. Pollitt *et al.*, 1985; Fisher-Hoch & Hughes, 1996; Ahmed & Pollitt, 2001) on valid and invalid sources of difficulty in examination questions is one example of how this approach has been applied.

If each year's test is constructed with these principles in mind, then all stakeholders will be able to be more confident in the assumption that consecutive years' tests are measuring the same construct, and hence that it is a more valid exercise to determine cut-scores at an equivalent standard, whether by statistical or other means.

Conclusion

Considering the issue of maintaining standards on consecutive years' tests as a measurement problem, which can be approached by using a statistical measurement

model, provides a rational basis for considering the effect of making questions more 'accessible'.

This approach shifts the focus from a semantic debate about the meaning of difficulty and accessibility, and unquantifiable assertions about the effect on motivation, to an attempt to construct a meaningful dimension or latent trait of ability, along which both questions and pupils can be unambiguously ordered. The 'standard' enshrined in each cut-score can be interpreted in psychometric terms as a particular point on the trait. Tests with easier questions will require a higher raw score to demonstrate the same level of ability as tests with harder questions. Analysing the misfit to the measurement model (both underfit and overfit) has the potential to draw attention to individual questions that may contain a different mix of dimensions to the majority of the questions. Such questions could be the ones where accessibility issues are causing a measurement problem.

Improving the accessibility of such questions may well have a beneficial effect on the quality of the test and will also make the test easier, by an amount that can be estimated from the model. The cut-scores will, and should, rise to account for the greater ease of the test.

References

Ahmed, A. & Pollitt, A. (2001) Improving the validity of contextualised questions, paper presented at the *British Educational Research Association Annual Conference*, University of Leeds, Leeds, September.

Andrich, D. (1989) Distinctions between assumptions and requirements in measurement in the social sciences, in: J. A. Keats *et al*. (Eds) *Mathematical and theoretical systems* (Oxford, Elsevier), 7–16.

Borsboom, D., Mellenbergh, G. J. & Van Heerden, J. (2003) The theoretical status of latent variables, *Psychological Review*, 110, 203–219.

Cresswell, M. J. (1996) Defining, setting and maintaining standards in curriculum-embedded examinations: judgmental and statistical approaches, in H. Goldstein & T. Lewis (Eds) *Assessment: problems, developments and statistical issues* (New York, Wiley).

Fisher-Hoch, H. & Hughes, S. (1996) What makes mathematics exam questions difficult? Paper presented at the *British Educational Research Association Annual Conference*, University of Lancaster, Lancaster, September.

Lord, F. M. (1952) *A theory of test scores* (New York, Psychometric Society).

Lord, F. M. (1980) *Applications of item response theory to practical testing problems* (Hillsdale, NJ, Lawrence Erlbaum).

Massey, A., Green, S., Dexter, T. & Hamnett, L. (2003) *Comparability of national tests over time: key stage test standards between 1996 and 2001* (London, QCA).

Masters, G. N. (1988) Item discrimination: when more is worse, *Journal of Educational Measurement*, 24(1), 15–29.

Michell, J. (1990) *An introduction to the logic of psychological measurement* (Hillsdale, NJ, Lawrence Erlbaum).

Petersen, N. S., Kolen, M. J. & Hoover, H. D. (1989) Scaling, norming and equating, in: R. L. Linn (Ed.) *Educational measurement* (3rd edn) (Phoenix, AZ, Oryx Press), 221–262.

Pollitt, A., Entwhistle, N., Hutchinson, C. & de Luca, C. (1985) *What makes exam questions difficult?* (Edinburgh, Scottish Academic Press).

Rasch, G. (1960) *Probabilistic models for some intelligence and attainment tests* (Copenhagen, Danish Institute for Educational Research).

Tomlinson, M. (2002) *Inquiry into A-level standards. Interim Report, September 27, 2002*. Available online at: http://www.dfes.gov.uk/docs/alevel_interim_report.htm (accessed 1 April 2005).

Wright, B. D. (1999) Fundamental measurement for psychology, in: S. E. Embretson & S. L. Hershberger (Eds) *The new rules of measurement: what every psychologist and educator should know* (Mahwah, NJ, LEA), 65 – 104.