This paper is to be published in the Royal Statistical Society Journal 'Significance', September 2004.

Measuring educational standards.

by

Harvey Goldstein

Institute of Education, University of London

h.goldstein@ioe.ac.uk

Introduction

'One of Britain's most senior examiners has revealed that GCSE maths results were fixed this year to mask the poorest marks by candidates for almost a decade' (Sunday Times, 14.9.2003).

'All independent evidence shows overall standards (in GCSE) to be rising since 1997. (David Milliband, Minister for school standards, Guardian, 18.9.2003).

Every Summer, when English public examination results appear there is a debate within the media about whether 'educational standards' are changing. The above remarks are typical; the first reflects the suspicions of those who believe in a professional conspiracy to prove that standards of performance are rising, and the second reflects a politician's determination to pretend that everything has been for the best since his party assumed power. In late Summer 2002 the chairman of the examinations oversight quango, the Qualifications and Curriculum Authority (QCA), was effectively sacked over allegations that he had sanctioned a downgrading of A level examination grades in order to maintain comparability with the previous year's results (Guardian, 27.09.2002).

Strikingly, but perhaps predictably, this annual 'debate' is long on rhetoric and short on evidence. It is also an obsession in only a relatively small number of educational systems in addition to that of England, and worldwide there seems to be very much less concern with this issue (Wolf, 2000). I shall argue that much of what passes for informed debate in the English media is both irrelevant and ill-informed, and in particular that there is no purely objective, statistical, means for measuring changes in 'standards'. For the reader who wishes to understand the nuances of the issue in more detail, the proceedings of a recent British Academy seminar (Goldstein and Heath, 2000) bring together existing understandings and I shall draw on some of these contributions. While this article focuses on public examinations, similar issues arise with regular government mandated Key Stage tests where the principal focus appears to be on measuring changes over time in order to achieve 'targets'. In this case, however, a key factor is the way in which such 'high stakes' testing regimes

encourage ‘teaching to the test’ which then distorts the desired comparisons. See Klein et al., (2000) for relevant evidence.

In the following sections I will outline the basic issues, discuss some of the methods that have been suggested for measuring changes over time and offer some views about how we might have a more useful debate about standards.

Changes over time

Start and Wells (1972) did a study of changing reading standards from the late 1940’s to the early 1960’s. They used results from repeated administration of the same test over this period and pointed out that the curriculum had changed and so had language use over this period and for these reasons they suggested that the test itself had become ‘harder’ so that apparent declines in test scores could not be viewed in any sense as a decline in standards of achievement. This duality of interpretation has long been recognised: in general, without making further assumptions, we cannot know, for example, whether the individuals taking a test or examination have in some sense become ‘better’ or whether the test has become ‘easier’ because the social, cultural or educational context has changed.

The key, therefore, to being able to make valid statements about changing standards over time resides in the reasonableness of the assumptions that are made. For example, over a very short period of time it may be perfectly acceptable to assume that ‘background’ factors such as curriculum are constant, so that the results from an application of a common test at two different time points does allow statements about changing standards of performance. This assumption lies behind many of the published tests from testing organisations. With high profile public examinations and regular government testing programmes, however, the use of the same testing instrument over time is clearly ruled out, so that other assumptions have to be made. In the next section I will briefly outline a statistical or ‘psychometric’ approach that has often been used for this purpose and in the following section I shall discuss the approach used by public examination boards.

Test equating

The idea behind the use of test equating procedures (see for example Holland and Rubin, 1982) assumes that one has two different tests administered at two different times, A and B. There are several variants but for simplicity I shall describe just two, the ‘common item’ procedure which underlies many practical schemes, and a sampling procedure.

In the first approach each test contains a small number of identical questions, say 15% of the total, a small enough number to avoid ‘detection’ but large enough to carry out a satisfactory equating. The idea is that these items are those that these are ‘invariant’, that is they can be assumed to retain a common ‘meaning’ for both time points while the other items are allowed to reflect changes in curriculum, general environment etc.

The common items are then used as a calibration set to create a common scale over all the items in the tests. This common scale is then used to report any changes. The actual procedures used to carry out the scaling vary in terms of the complexity of modelling used, but typically some kind of binary factor analysis model is used, often referred to as 'item response theory'. The problem, however, is twofold. First it is necessary to make the invariance assumption for the common items and this, inevitably, is a matter for judgement which may not be universally shared. Secondly, even if such an assumption is accepted, because the non-common items are allowed to reflect background changes, the relationship between the common item set and the non common items can be expected to vary across the tests; yet it is necessary to assume that this relationship is constant. This second assumption is therefore contestable and also a matter for judgement. An interesting example of these problems arose with the US National Assessment of Educational Progress where there was a very large and unexpected drop in test scores over a 2-year period in the 1980's. A large scale evaluation concluded that essentially the common item equating procedure was unreliable, for a variety of reasons including their juxtaposition with different surrounding items in the two separate instruments (Beaton and Zwick, 1990).

In the second approach the idea is that a very large 'bank' or 'pool' of items is selected and for each test a random, possibly stratified, sample is selected for use. This means that, apart from sampling error, a common scale does exist and can be used for inference. Such procedures are often referred to as 'item banking', although that term is also used in other contexts. The difficulty is that the pool has to be selected *before* any of the tests are administered and it cannot be known in advance which items may become outdated and hence become 'harder' etc. Thus, again, assumptions about test item behaviour have to be made.

In both these cases I am not necessarily arguing that these approaches are pointless, nor that test equating is not useful in other situations. Rather, I am suggesting that they are not simple 'objective' devices for solving the problem but in fact involve important, and crucial, value judgements that may or may not find consensus among interested parties. One of the unfortunate aspects of much of the literature on equating is that this need to exercise value judgements is rarely stated (see Goldstein, 1994, for a more detailed critique).

Setting standards

The procedures used by UK public examination boards in attempts to maintain standards of grading over time are described in detail by Cresswell (2000) and constitute one of a number of procedures generally referred to as standard setting methods (Wiliam, 1996). In some circumstances, at least in the past, a simple 'norm-referencing' was carried out whereby a constant distribution over grades was maintained from year to year. With changing examinations, greater competition among examination groups and external demands, this is generally no longer the case, but there is nevertheless an attempt to maintain strong statistical relationships between grade distributions in successive years.

Public examination questions are set by experienced examiners who attempt to match 'difficulty' across years in the light of changing curricula. Nevertheless, when marks become available, it is typically the case that some questions appear to be harder or easier than anticipated. This appearance, of course, is a matter of judgement but this judgement then informs decisions about where to draw grade 'boundaries' in relation to marks. Thus, a boundary presumably would not be drawn if it resulted in a very large change in the grade distribution, *unless* there was felt to be a very good reason why, say, performance had deteriorated. Such a reason might of course be present – for example a disruption in schooling due to a natural or human disaster, but I am not aware that such a reason has ever been used. The point is that examiners are making judgements about relationships, this time in terms of the stability of grades. In other cases, where large changes in numbers or types of candidates occur, they will need to make further assumptions about the nature of the candidates sitting the examination. Thus, the first quotation at the start of this article may reflect a perfectly reasonable attempt to adjust for curriculum changes, and in fact the examiner quoted did claim subsequently that the newspaper had distorted his remarks (BBC education web site, 15/09/03). The point here is that, very often, an assumption of population stability is built into the setting of performance standards themselves so that the resulting outcomes cannot be used directly to make inferences about those standards.

Conclusions

It does seem quite natural to wish to measure changes in educational performance over time. After all, we can measure whether populations are getting taller of heavier over time and such indicators can be very useful. Education characteristics, and mental characteristics more generally however, are different in that the devices used to make measurements themselves need to be defined within an existing context. This may be a school curriculum, a cultural background that affects language, a changing tradition of teaching and so on. It is this contextual specificity that creates the problems and tends to rule out the possibility of any generally agreed 'objective' common measurement scale.

Having said that, however, there is a role for informed judgement. Thus, it will often be possible to infer that certain 'standards', say in literacy, have changed over a long period of time by appealing to indicators such as newspaper readership or textbook content. Assumptions will be present here too, but may be generally acceptable. The more difficult judgements are those that need to be made over short time periods where relatively subtle changes will have taken place.

Rather than concentrating on measuring absolute changes over time, it would be more useful to focus on relative changes. Thus, for example, changing gender or social class differences are both more interesting and more amenable to study. Naturally, interpretations are still problematical; gender differences may have changed as a result of changing examination formats, say from essay type questions to multiple choice. Such possible explanations, however, can be further explored and may suggest fruitful research topics (see for example, Elwood and Comber, 1996). Likewise, it may be informative to model how the variability in scores or grades changes over

time. What is clear, however, is that there is little to be gained and much to be deplored in the way that current public debates have centered on sensational, but ultimately rather sterile, discussions about absolute standards.

References

Beaton, A. E. and Zwick, R. (1990). *Disentangling the NAEP 1985-1986 reading anomaly*. Princeton, Educational Testing Service:

Cresswell, M. (2000). The role of public examinations in defining and monitoring standards. *Educational Standards.* H. Goldstein and A. Heath (Eds.). Oxford, Oxford University Press.

Elwood, J. and Comber, C. (1996). *gender differences in examinations at 18+*. London, Institute of Education.

Goldstein, H. (1994). Recontextualising mental measurement. *Educational Measurement: Issues and Practice* **13**: 16-19.

H. Goldstein, and A., Heath., (Eds.) (2000). *Educational Standards*. Oxford, Oxford University Press.

Holland, P. W. and Rubin, D. B. (1982). *Test Equating*. New York, Academic Press:

Klein, S. P., Hamilton, L. S., McCaffrey, D. F. and Stecher, B. M. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives* **8**: 1-21.

Wiliam, D. (1996). Meanings and consequences in standard setting. *Assessment in Education* **3**: 265-286.

Wolf, A. (2000). A comparative perspective on educational standards. *Educational Standards.* H. Goldstein and A. Heath (Eds.). Oxford, Oxford University Press.