Measurement in education.

Ian Plewis

Measuring pupils' educational attainments and achievements is not as simple as measuring their heights and weights, even though some public pronouncements would suggest that it is. In this article, I consider four issues, each of which links to public debates in education. These issues are:

1) Comparisons between different measures at one point in time.
2) Comparisons over time, i.e. between cohorts.
3) International comparisons of attainment.
4) Developmental or longitudinal changes with age.

My conclusion is that, rather than trying to answer the unanswerable, statisticians should try to persuade policy makers and others to reformulate their questions for each of these four situations.

1. Comparing different measures at a single point in time.

Here we are concerned with questions such as performance in different A-levels, comparing levels across National Curriculum subjects, and comparing levels across Attainment Targets within National Curriculum subjects. If we were to use a biological analogy then the first two are like comparing upper and lower limb strength, which, because we use our arms and legs for different activities, is not terribly useful. However, the third is like comparing left and right eye vision which can be useful, especially when you want a prescription for glasses.

Facts:

1. Using a measure of 'relative ratings' (described by FitzGibbon and Vincent, 1994), Physics appears to be the most difficult A-level. (Note however that, in 1995, three times as many candidates got an A in Physics compared with Business Studies.)

2. In the 1996 Key Stage One national assessments, 6% scored above level two in Writing, 15% in Number.

3. In the 1995 Key Stage One national assessments, 20% scored below level two and 15% above level two in Number whereas, for Handling Data, the corresponding percentages were 25% and 10%.

There are a number of statistical issues which we must address when trying to interpret these facts. As far as comparing A-levels goes, we need to think about sampling and self-selection because different students choose different subjects and different mixes of subjects, we need to consider the effects of mixes across examination boards, and the effects of student effort and motivation (Goldstein and Cresswell, 1996). When comparing A-levels and when comparing national assessments across the core subjects in the National Curriculum, we need to consider whether marking is more stringent for some subjects than for others. When comparing results for different parts of the mathematics curriculum, we need to consider developmental issues - is 'handling data' inherently more difficult for young children than 'number'? And for all three of the above facts, we must think about the quality of teaching and how it might vary across subjects

Rather than dwelling on what might be labelled the 'media studies syndrome' that some A-levels are soft options, one more interesting approach to understanding and interpreting these educational facts, would be to develop measures of teaching quality across the curriculum which might help us to account for some of the observed differences.

2. Comparing performance across time - between cohort comparisons.

Here the questions are about A-level standards across time, and reading standards across time, and the biological analogy is looking at between cohort differences in height.

Facts:

1. 76% of candidates got grades A to E at A-level in 1989, 84% in 1995; 11.4% of candidates got grade A in A-level in 1989, 15.8% in 1995.

2. 12% of the 1958 birth cohort, 20% of the 1978 birth cohort got at least 2 A-levels.
(OFSTED, 1996)

The statistical issues to be addressed here include gender mix (an increasing proportion of girls are taking A-level) and gender dispersion (boys' results are more variable than girls'), exam board mix, subject mix, institutional mix (changes in the proportions taking A-levels in state and independent schools, Sixth Form colleges and FE colleges), changes in the form of examinations, and changes in the information available to teachers about syllabi and what is expected of candidates.

Rather than hanging on to a nostalgic view of how A-levels used to be really difficult and now 'nobody knows what a gerund is any longer', and not being prepared to recognise that just as young people are getting taller, perhaps they are also getting cleverer, a more interesting approach would be to ask whether the match between A-level curricula on the one hand and HE courses and employers' requirements on the other, could be better than it is (and, perhaps, was twenty years ago).

3. International comparisons.

Here we are concerned with between country differences in mathematics and science performance as revealed by, for example, the recent TIMSS (Third International Mathematics and Science Study) research. (Keys et al., 1996). A biological analogy is that Swedes are, on average, taller than the Japanese.

Fact:

The median score at age 13 for mathematics for Japan is 572 whereas for Sweden it is 497.

The statistical issues here include those to do with sampling, variability within and between schools and how these might affect inferences, and curriculum differences sometimes known as the 'opportunity to learn'.

Lamenting that pupils in Singapore are so much better at mental arithmetic than English students are, and hence falling into the 'ecological fallacy' trap of supposing that teaching methods in Singapore must therefore be better than teaching methods in England, is not helpful any more than it is to suppose that the Swedish diet is better than the Japanese diet. A more interesting approach is to recognise that TIMSS-type studies are not an end in themselves but a stimulus to further research. For example, can we link differences in teaching methods, classroom organisation and so on to differences in attainment and progress.

4. Describing differences across age - longitudinal or developmental change.

This is a more technical issue, where we might be thinking more about patterns of change across the ten point scale used in national assessment. The biological analogy is the usefulness of looking at growth norms from ages x to y rather than height norms at ages x and y.

The statistical issues concern the validity of the ten point scale for national assessment. Here the lack of research into the measurement properties of the scale is scandalous, and makes an extraordinary contrast with the emphasis placed on measuring school performance with league tables.

A more useful approach would be to focus on relative change - differences between groups - rather than on absolute change and the shape of growth curves.

Some of the facts set out earlier lead to questions and puzzles and statisticians have a lot to contribute by way of unravelling these puzzles. However, rather than engaging in debates about education, the terms of which are characterised by a longing for the past combined with a wish to denigrate teachers at every opportunity, perhaps the most important contribution statisticians can make is to suggest that the way questions are often posed in public debates is not necessarily the best way, and that there are other, potentially more useful questions which should be asked. Moreover, these questions might turn out to be easier to answer.

References

FitzGibbon, C.T. and Vincent, L. (1994) *Candidates' Performance in Public Examinations in Mathematics and Science*. London: SCAA.

Goldstein, H. and Cresswell, M. (1996) The Comparability of Different Subjects in Public Examinations: a theoretical and practical critique. *Oxford Review of Education, 22*, 435 - 442.

Keys, W., Harris, S. and Fernandes, C. (1996) *Third International Mathematics and Science Study - First National Report; Part 1*. Slough: NFER.

OFSTED (1996) *Standards in Public Examinations 1975 to 1995*. London: OFSTED and SCAA.

Author: *Ian Plewis, Thomas Coram Research Unit, 27/28 Woburn Square, WC1H 0AA. email: tesp102@ioe.ac.uk.* This is a written version of a talk given to the ESRC Analysis of Large and Complex Datasets conference, Warwick April 1997.