to expose this situation, but one should be alive to the danger that if one rocks the boat too much there will be pressure for standardising syllabuses, and prescribing the criteria that should be tested in any English (or Mathematics etc.) examination. There are already pressures for this from the Standing Conference on University Entrance at A-level; and the Waddell Report, reinforced by the previous Government's White Paper and subsequent discussions about the establishment of a National Co-ordinating Body, moves in this direction for a common system at 16+.

For a school-leaving certificate as a replacement for or supplement to public examinations, one faces a dilemma: should one pander to the genuine, if somewhat (but only somewhat) misguided, desire for comparability by setting up inevitably elaborate procedures for moderation, validation and accreditation and give labels, either in words or in grades, that are designed to have some common currency between schools (but warning of the limitations of the common-ness of the currency) or should one allow each school to develop its own reporting procedures? The latter invites employers, universities and colleges to take much more initiative in devising their own selection procedures: in many senses this would be thoroughly constructive, but it does invite 'league tabling' of schools, nepotism and all those bad effects that public examinations were originally devised (first in China 3,000 years ago, and in the 19th Century in this country) to overcome. It also implies a boom in (often inappropriate) psychological and educational testing by employers with possible unfortunate backwash into the schools, let alone the creation of problems for children on the milkround of job seeking. How can one avoid leaping from the frying pan into the fire?

---

1 This paper was originally written in the summer of 1978 and delivered at a PRISE Conference on Assessment held on 31 March, 1979, at Oxford. It has been slightly amended for publication.

2 For a development of these arguments and more technical detail about the problems of investigating comparability, see Comparability in GCE published by the JMB on behalf of the GCE Boards (May 1978) and Comparability of Standards in Public Examinations: Problems and Possibilities prepared by the Schools Council Forum on Comparability and to be published shortly by the Schools Council.

3 See, for example, The Reliability of Examinations at 16+ by A.S. Willmott and D.L. Nuttall (Macmillan, 1975).

Changing Educational Standards : A Fruitless Search

by Professor Harvey Goldstein, Head of Department of Statistics and Computing, Institute of Education, London University.

To make comparisons across time, to chart the progress of institutions or societies, is such a common activity that one rarely finds any serious attempt to question either its usefulness or its feasibility. Many individuals exert a great deal of effort, for example, to devise standard of living or price indices to monitor aspects of economic change. Others calculate mortality rates or estimate the average heights of children in order to monitor progress in health. In both cases, but especially the former, there may be arguments about the most useful technique to use, but little dispute about the desirability of making comparisons over time or about the possibility in principle of being able to do so. For educational attainments, likewise, the notion of making comparisons across time is deeply embedded both within the examination system and with regard to standardized achievement testing. It is only quite recently that there has been any suggestion that comparability of exam performances across time may have inherent problems which defy purely technical solutions. Such a possibility, however, does not seem to have been taken seriously in current discussions of across time comparisons of 'standards' based on achievement tests. In this article I shall air certain difficulties about such acrosstime comparisons, and suggest that we may have been expecting answers to the wrong questions.

In the so-called 'Great Debate' in education, the contribution of the Government by way of the Assessment of Performance Unit (APU), and the increasing use by LEA's of standardized tests of language and mathematics are now familiar and have become incorporated within the ubiquitous 'Educational Accountability' scene. A central motivation for this activity, and an explicit aim of the APU, is to make useful statements about changes in standards of performance over time. Such statements might, for example, concern the changing mathematical competence of school leavers. Thus, an engineering employer might feel that the mathematical attainments of the school leavers he employs compare unfavourably with those of 20 years ago. (I hasten to add that any examples I use are chosen to illustrate points and not to score them). Such a statement, however, says little about the mathematical attainments of school leavers in general, since the type of leaver entering engineering may well have changed over time with possibly more of the mathematically able now going to higher and further education, or other kinds of employment. By the same token, statements about the achievements of university entrants say little about achievements among 18-year-old school leavers. Moreover, even if we did have evidence about changing attainments for children in general, it would not necessarily follow that the cause of any changes lay in the schools. Falling attainments, for example, might be due to changing environmental factors which are themselves related to mental functioning. The evidence needed to connect the school system with change in attainments is nearly always lacking and indeed it is difficult to see how one might obtain such evidence, since school curricula and organisation are changing at exactly the same time as the wider society and environment are changing and there seems little hope of disentangling these factors. Of course, specific aspects of a curriculum or types of school organisation can be compared in a research project, but this is not the same thing as separating out influences which are historically confounded. Where it is possible to make comparisons of changes for different sub-groups of the population, for example, those in the north of the country compared to those in the south, then such relative changes might be more informative. Here again, though, without further specific research there is little possibility of ascribing anything directly to the effect of the school system. Finally, there is the real difficulty of trying to

measure 'the same thing' over time. Let me elaborate on this.

One of the most ambitious recent attempts to assess changes in educational achievement was the work of Start and Wells (1972) who studied a series of reading tests from 1948 to 1971. They encountered all kinds of difficulties, but the major and insoluble one arose when it became clear that there was no single reading test which it was appropriate to use over this time span. (There is, of course, the other major difficulty that many experts would disagree over just what was a good reading test at any time). Two tests, the Watts-Vernon and the NS6, were used, but both contained some items which were less relevant to children at later times than at earlier times, thus appearing to become 'harder' over time. For example, the term 'mannequin parade' which occurs in the NS6 and may have been familiar in the 1950's, would be much less familiar by 1970. Thus, any apparent changes in average test scores, for example, might simply be due to a test becoming outdated and thereby more difficult without necessarily reflecting 'lower' achievement. We could only say that, with respect to the test, achievement had deteriorated. This would not be a very useful statement to make, however, if the test itself was felt to be no longer appropriate. Of course, the reasons for test items becoming dated might lie in the curriculum, teaching practices, or in society at large, but it is generally agreed that they do occur. Furthermore, they may even occur over relatively short periods of time when, say, rapid language changes are taking place or when cheap electronic technology is influencing numeracy skills. When use of the same test over time is not valid, what other options are open?

One possibility which has been canvassed is the use of different test instruments which are calibrated against each other so that scores on a later test can be converted to those on an earlier one. Unfortunately, this procedure makes some strong assumptions about the invariance of the calibration relationships over time which will almost certainly not be true as the earlier test becomes outdated. In fact, all such calibration devices which depend on the stability of test or item relationships over time seem to be inherently unworkable. A more detailed discussion of the difficulties in using a sophisticated procedure of this kind known as 'item banking' is given elsewhere (Goldstein, 1979a, 1979b). The other possibility is to construct carefully a new test which contains only those items considered to be applicable fully over the time scale of interest. The difficulty about this proposal is that it involves making predictions about the future which are notoriously risky. For example, the advent of cheap pocket calculators has changed attitudes to arithmetic and it is difficult to see how a test of 'numeracy' devised in 1970, could have been fully relevant also to 1980. One might be able to conclude that, say, the ability to carry out rote arithmetic had changed, but such information may not be very useful. If we wished to devise a test of numeracy in 1980, we would need to recognise the advent of the new technology in the test items.

This discussion leads to some general conclusions. First and foremost, it seems that the search for absolute comparisons of achievement over time is a fruitless enterprise. Secondly, the pursuit of such comparisons is wasteful of resources which could otherwise be devoted to devising new test instruments which recognise the true nature of educational innovation and which are designed to be appropriate for a particular context. If we are satisfied that a test is appropriate and relevant at a particular time then we may use it to compare children in different environments etc. Thus, we may wish to know whether boys and girls have different 'numeracy' skills in 1980 and whether any difference appears to be greater than any difference which existed in 1970. The tests would be different at the two occasions, but would have been devised in order to try to measure what is meant by 'numeracy' in an appropriate way at each occasion.

In conclusion, I would like to reiterate that I believe we should drop the idea that it is possible to talk of absolute changes in educational attainments over time. It would be more fruitful for test constructors to pay attention to producing tests which are as up-to-date and as relevant as possible to a changing educational system and to the wider environment, rather than to perpetuate outdated instruments or to pursue dubious calibration procedures on the grounds of comparability over time.

Acknowledgements

I am grateful to Desmond Nuttall, Ian Plewis and Bob Wood for their useful comments.

References

Start, K.B. and Wells, B.K. (1972)
The Trend of Reading Standards, N.F.E.R., Slough.

Goldstein, H. (1979a)
'Objective measurement' and the mystification of educational assessment in Forum, September 1979.

Goldstein, H. (1979b)
Consequences of using the Rasch Model for educational assessment in British Educational Research Journal, Vol.5, No.2.

Re-organisation and In-Service Training in one area of Surrey

by B.J. Canton, Area Inspector, Surrey.

The Farnham and Ash area of South West Surrey was re-organised in 1973 along comprehensive lines. Where there had previously been two single sex grammar schools, four secondary modern schools and a number of primary schools, the pattern was now changed as follows; one co-educational, open access VI Form College based upon the boys' grammar school; four comprehensive 12-16 schools; eleven middle schools for the 8-12 age range; 19/20 first schools for the 5-8 age range.

The changing structure, of itself, created the need for considerable rethinking on a wide range of matters, some practical and immediate, some more theoretical and far reaching. To facilitate such discussion a Steering