'educational panics' based on rather flimsy evidence. Goldstein (1983) reviews some of the methodological issues. The problems associated with setting targets for attainment at different ages are raised by Plewis and Goldstein (1997). The logic of making the kinds of comparisons illustrated by Figure 31.1 are set out by Plewis (1985). The importance of using a multilevel approach for the analysis of all kinds of educational data, and the link between multilevel modelling and the ecological fallacy (Figure 31.2), are set out by Plewis (1997). The 'opportunity to learn' idea is described by Plewis and Veltman (1996).

Statistics in Society. D.
Dorling & S. Simpson (eds.)
Arnold.

32

Performance Indicators in Education

Harvey Goldstein

Knowing about uncertainty in performance indicators prevents their misuse

The major impetus behind the extensive collection and publication of performance indicators for schools in England and Wales came with the Education Reform Act of 1988. Prior to that it had been possible to collect public examination data on schools, but although these were sometimes published locally, such data were not the subject of systematic rankings or 'league tables'. Since 1988 the collection and publication of average results for schools has become routine and extends all the way from test scores at the age of 7 to A-level examination results at the end of schooling.

The principal official support for the 1988 legislation was a report produced by a task force (Task Group on Assessment and Testing, 1987). This report sketched out a national framework for educational assessment and testing, and made particular recommendations on the publication of test scores and examination results. It recommended that aggregated (mean) scores for each school should be published, but only 'as part of a broader report by that school of its work as a whole' (para. 132). That broader report was to include aggregate data about the nature of the school catchment area, and other 'contextual' data the school considered relevant to interpreting its results. The TGAT report briefly discussed the issue of whether school results should be adjusted for factors outside its control and argued that it 'would be liable to lead to complacency if results were adjusted, and to misinterpretation if they were not' (para. 133). It went on to conclude that 'results should not be adjusted for socio-economic background but the [school] report should include a general statement . . . of the nature and possible size of socio-economic and other influences which could affect schools in the area'.

Two important issues arise immediately from these highly ambiguous, and somewhat contradictory, statements. First, they betray a considerable level of ignorance about the purpose and nature of 'adjustments', and I will return to this in the next section. Second, they neatly illustrate how a statistical issue, namely how to construct and interpret a statistical model explaining school performance, can be dealt with by the use of plausible, but fairly meaningless, language. In my final section I shall return to this issue of how politicians and others are able to appropriate aspects of mathematical and statistical terminology in order to avoid

confronting the complexities of real data and the need to be clear about the caveats and limitations of such data. I should of course make it plain that such appropriations and consequent distortions are not necessarily deliberate and may often be done for what are perceived of as the best of motives; my concern is more with the fact that we inhabit a culture which tolerates such incoherence.

In 1992 the first national league tables of public examination results were compiled by the then Conservative government and published in the main national newspapers, with annual reports thereafter. This policy has now been extended to test scores at 7 and 11 years, and is a policy endorsed and extended by the Labour government elected in 1997. These tables, with a minimum of contextual information, have had a profound effect on schools (Ball and Gewirtz, 1996) which see themselves as competing for 'customers' in an educational marketplace. The Citizen's Charter (UK Government, 1991) formalized this by extending the TGAT recommendations to require schools and colleges to 'publish their annual public exam results in a common format' so as to provide 'easier comparison of results between schools'. The official justification was 'to help parents choose a school'. Interestingly, as a result of widespread concern, in 1995 the government officially endorsed the principle of contextualisation when it recommended the use of 'value added' comparisons among schools (Department for Education, 1995), and subsequently a report was produced by a government quango, the Schools Curriculum and Assessment Authority (1997), which sought to operationalise this. Nevertheless, this implicit admission of inadequacy has not affected the practice of publishing 'raw' or unadjusted comparisons. For reasons I will go into below, even if implemented, the use of value-added data would not resolve most of the underlying problems.

Performance indicators and public accountability

The basic political justification for the publication of league tables, whether in education, health or social services, is that the 'public' has a right to know something about the performance of publicly funded institutions. Such knowledge can then be used, for example by government itself, in deciding whether to take action against institutions perceived to be 'failing', or by parents, say, in choosing a school for their child. For information of this kind to be useful it clearly has to meet certain quality standards. It should be reliable enough to make useful distinctions among institutions and it should be valid in the sense that it really does reflect the qualities claimed for it, whether these are standards of educational delivery, health care or social service provision. In the remainder of this chapter I shall show that current performance indicators fail on both these counts: they are unreliable and they distort the underlying reality.

It is of some interest, given the shaky intellectual foundations of current performance indicators, to ask why they have been promoted with such vigour by successive governments. This is not the place for a detailed analysis, but a few observations may be useful. First, there has been a great centralising tendency of governments since the late 1970s. To some extent obscured by the free-market rhetoric of Thatcherism, there has been, nevertheless, especially in education, a transfer of power from local to central government and at the same time a transfer from professionals, i.e. teachers, to government or quango employees and also

to other bodies such as governors. Performance indicators have served to control both schools and the teachers within them as an external yardstick which has forced adherence to a nationally imposed curriculum and testing regime.

Second, and closely related to centralisation, governments have often been suspicious of educational professionals as potentially 'subversive', users of unfamiliar language and with a perceived great influence over future citizens. The reaction has taken the form of attempting to simplify educational debates. Thus, the 1997 Labour administration has made much of 'standards', by which it means achievements on readily understandable tests and examinations. It appears to be relatively uninterested in the real complexities surrounding teaching and learning. The rhetoric of 'standards' has even extended to the setting of 'targets' for schools to achieve certain test scores several years in advance with little concern for whether this is really feasible (*see* Plewis and Goldstein, 1997, for a critique).

Whatever the reasons for the politicians' interest in performance indexes, it does seem fairly clear that these satisfy a deeply felt need and this therefore makes a proper debate about their status very important.

How should we compare schools?

In this section I want to take some examples to show how problematic this issue of school, or any other institutional, comparisons really are. I shall draw heavily on a technical review (Goldstein and Spiegelhalter, 1996) which sets out the issues in some detail, as well as other research which has been extensively replicated.

The principal argument against examination league tables is that the performance of a school is determined largely by the pre-existing achievements of the students when they enter it. Since schools differ markedly in this respect – for example, some schools are highly selective – it is impossible to judge the quality of the education *within a school* solely in terms of final 'outputs'. There are also, however, problems which apply to 'value-added' tables, and I will show how initial expectations that these could provide a more sensitive indicator of school performance have failed to materialise. Furthermore, attempts to adjust 'raw' results using average socio-economic background or even average intake scores of students are inadequate. Thus Woodhouse and Goldstein (1989) showed that attempts to do this resulted in highly unstable rankings, and that small and essentially arbitrary decisions about how to formulate the statistical models led to very different conclusions. Thus, the suggestion in the appendix to the 1997 education White Paper (Department for Education and Employment, 1997a) that an adequate adjustment can be made using the percentage of pupils in a school having free school meals, is invalid. Proper adjustments can only be made, at the very least, if *individual student* intake scores are taken into account.

Nevertheless, even if an adjustment can be made using individual student data, several difficulties remain. The first problem is that typically only a single figure is reported, such as the overall percentage of high GCSE grades. Yet schools may be 'differentially effective'. Thus, for example, two schools may perform equally well on average but one may have poor performance in mathematics and good performance in English and the other vice versa. Likewise, where value-added tables are concerned, some schools may exhibit relatively good performance for initially (on intake) poorly achieving students and produce relatively weak

performance for initially highly achieving students, and vice versa for another school (Goldstein *et al.*, 1993). A second problem, with both 'raw' and value-added tables, is that the percentages or scores produced for each school typically have a large margin of error or 'uncertainty' associated with them. This problem is even more acute when individual subjects or departments within schools are the focus of interest, since the sometimes small numbers of students involved mean that very little can be said about any individual department's performance with reasonable accuracy. In the extreme case, for some A-level subjects there may be only two or three students involved, and any generalisation, even over a number of years, from such small numbers is extremely hazardous.

Figure 32.1 illustrates this general problem. It is taken from a survey of some 400 schools and colleges with A-level results where value-added scores are calculated by adjusting for the GCSE performance of the candidates (Goldstein and Thomas, 1996). Each bar corresponds to a value-added score for a school. The fewer the number of students in a school the longer is the bar and the more uncertainty is attached to the value-added score. The bars are ordered from left to right on the school's value-added score (the mid-point of the bar). The bar lengths are chosen in such a way that two schools can be judged as being reliably separated only if their bars do not overlap. Thus, for example, the bars from the extreme left-hand school and the extreme right-hand school do not overlap and we may conclude that there is a real underlying difference between their value-added scores. In this figure, however, for some three-quarters of all possible comparisons of pairs of institutions, it is not possible to make such a separation. In other words, finely graded value-added comparisons are of limited value since in most cases we will find no difference.

Another problem with all of these tables is that they refer inevitably to a cohort of students who began their education at those institutions many years earlier. Thus, for example, GCSE results published for a particular year refer to a cohort starting at their secondary schools some 5 years previously: given that schools can change markedly over time, there will be some uncertainty over the use of those results to predict the performance of future cohorts.

A further problem arises from recent research (Goldstein and Sammons, 1997a) which shows that the primary school attended by a child exerts an important influence on GCSE performance and that this should therefore be taken into account when producing value-added tables. Also, there are other factors, such as sex, ethnic origin and social class background, all of which are known to be associated with performance and progress throughout secondary schooling and which therefore will affect the interpretation of any rankings. Finally, there are several practical problems associated with producing any kind of performance tables based upon test or examination results, perhaps the most important being that during the course of a period of schooling, say from 11 to 16 years, many students will change schools. To ignore such students is likely to induce considerable biases into any comparisons, yet to include them properly would require enormous efforts at tracing them and recording their results.

Taking all these caveats together we can see that attempts to rank educational institutions are fraught with difficulty. Even with extensive and good-quality information, there are some inherent limitations which preclude the use of rankings other than as initial *screening instruments* to isolate possibly high- or low-achieving institutions or departments which can then be further investigated;

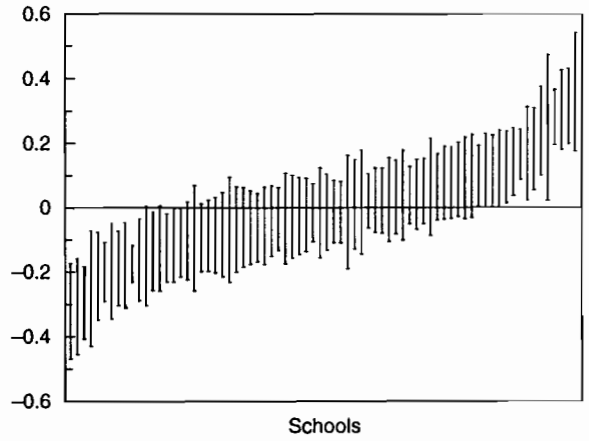

Figure 32.1 A-level scores: pairwise (95 per cent) uncertainty levels for a random sample of schools and colleges for students in the middle (50 per cent) GCSE score band. The data refer to the 50 per cent group of students who have an average GCSE overall subject score in the middle range. Schools are to be judged as statistically significantly different at the 5 per cent level only if they have intervals which do not overlap

bearing in mind that the information is historical. These caveats apply particularly to the public presentation of comparative tables. For internal 'school improvement' purposes, however, schools can often extract useful information from knowing where they are ranked among schools with similar characteristics, especially where detailed information about differential effectiveness is available. For such information to be useful it is essential that it be presented with all the necessary caveats and that it remain confidential to the school so that it can be properly evaluated in context. Several local education authorities are now beginning to develop such schemes. In Hampshire, for example, over 100 primary schools are taking part in a system where value-added results for different curriculum subjects and for different kinds of pupils are provided for each school in terms of progress between school entry and age 7 and between age 7 and age 11. The schools use these to help themselves to understand their strengths and weaknesses in comparison with other schools in the local authority, bearing in mind the inherent limitations of such analyses. (Chapter 35 examines other developments in Hampshire's education.)

A detailed discussion of the ethics of performance indicators and some suggested guidelines is given by Goldstein and Myers (1996), who also discuss the issue of how *particular* measures, such as examination results, have come to assume a dominant role in evaluating the performance of schools. They point out in particular that if comparisons among schools are to be attempted, it is very important to provide users with careful descriptions of all the limitations.

Social manipulation of statistical information

I have already referred to the fact that statistical information can become ambiguous and incoherent when its terminology is used for political or similar purposes.

In fact, this is a much more serious problem than that of mere abuse for particular purposes. It symptomises a cultural attitude towards quantitative information which informs discussion of social issues at all levels.

Statistical analysis has two key components: first, the modelling or summarisation of a set of data in terms of a small number of 'parameters', for example an average examination score; and second, a statement about the precision of the summary measures. Thus, for example, it is a common practice when reporting some opinion poll results to quote a percentage in favour of a course of action, plus or minus a 'margin of error' due to sampling fluctuations. In the previous section this error was expressed in terms of confidence intervals. We saw how the use of such intervals prevented any precise comparisons among institutions. Apart from their use as crude screening devices or as additional pieces of information for use by schools for improvement purposes, league tables, of whatever kind, have severe limitations.

Absorbing uncertainty when making judgements appears to create severe problems and is a major difficulty in conveying statistical results. This uncertainty may arise from sampling variation, as I have already discussed, or it may be due to difficulties in finding reliable measurements, or to lack of response from certain pupils in schools, etc. All of this is familiar territory to experienced statisticians, and forms a part of most research reports. Yet all too often such caveats are ignored. It is as if there is an assumption that numerical results *must* accurately reflect reality. The common view of mathematics and mathematically based science is that it deals only in those things which have accurate numerical representations. It is particularly unfortunate that this misunderstanding often accords with the common demand from politicians for justification of a position on the basis of 'hard facts', which can tolerate no uncertainty – and this makes any change much more difficult to envisage. What seems to be required is a cultural shift in attitude along with a positive attempt to incorporate a fuller understanding of statistical information and uncertainty into education at all levels.

It would not be too difficult to set out guidelines for the reporting of such things as performance indicators (Goldstein and Myers, 1996), and perhaps the single most important innovation would be the *mandatory* inclusion of uncertainty estimates. There are very few instances where this would not be possible. It can be readily justified in terms of freedom of information, on the grounds that such uncertainty estimates are a key component of any publication and that it is misleading to withhold them. In a democratic society there would seem to be little excuse for refusing to provide citizens with the caveats which are implicitly attached to public indicators of performance. A requirement to make such information prominently available could be incorporated into any new freedom of information legislation, and this would constitute a very important step towards mitigating some of the more harmful effects of performance indicator publication that we have seen.

33

Can Trends in Reading Standards Be Measured?

Pauline Davis

The use of the National Curriculum's Standard Assessment Tasks (SATs) for assessing trends in reading standards

> School Standards Minister Estelle Morris today welcomed improved 1997 National Curriculum test results for 7, 11 and 14 year olds in English .... These results show our continuing highlighting of the importance of literacy ... – and primary school homework – is clearly having a helpful effect in the classroom.
>
> (Department for Education and Employment, 1997d)

How can we know if such a claim is true? Are children's literacy standards improving or declining? Are children from all sections of British society exhibiting a common trend, or are some groups of children showing an improvement in standards while others are showing a decline? Why are the National Curriculum tests assessed as they are? Why do the test results take the form they take and why did they come into existence when they did? This chapter considers reading standards in conjunction with the Standard Assessment Tasks (SATs) of the National Curriculum of England and Wales in order to explore some possible answers to these questions, and in particular to consider the detection of variations in national reading standards over time (an issue raised for study in Chapter 30).

National Curriculum

The reasons for the introduction of the National Curriculum and SATs can be usefully described in a historical context, but a detailed review is beyond the scope of this chapter (*see* Chitty, 1989). However, it was in the 1980s that the supporters of policies typified by those of the Thatcher government commonly became known as the 'New Right'. The New Right rejected the social democratic-type policies, which had previously been accepted by Conservative and Labour governments alike since the introduction of the modern welfare state. 'New Right' policies were summarised by Gamble (1988), as 'free economy/strong state'. This bedding down of neo-liberalism, advocating freedom of choice, the individual, minimal government intervention and *laissez-faire* economics, seems to lie