194

New Economy

Relegate the leagues

Data from performance tables is crude and often misleading

HARVEY GOLDSTEIN

Professor of statistical methods, University of London, Institute of Education

It is now common to use student achievements to judge the quality of education in individual institutions. Examination results from GCSE and A levels, test scores from National Curriculum, and internationally collected test scores are all used to make comparisons among schools and other educational institutions. The justification for this rests upon the assumption that schools are very largely responsible for the skills and understandings which students acquire and that the examinations and tests are designed to measure.

However, educational institutions such as schools have other roles: social and community ones for example. They also have a responsibility for encouraging learning across a much wider range of areas than can reasonably be tested. Any judgment based upon the measurement of only a partial set of features has to be recognised as incomplete. My intention is to examine the usefulness and reliability of the evidence obtained from testing children for making judgments about institutional differences. First, however, it will be useful to say something about the purposes of assessment.

> *"Attempts to make use of an assessment, designed for one kind of activity, for a different purpose may be inappropriate and may lead to undesirable consequences."*

It is possible to isolate three principal purposes (Goldstein 1991).

The purposes of assessment

Assessment, sometimes informal, sometimes using formal procedures such as standardised tests, is used for diagnosis and the enhancement of learning. This is done by attempting to understand a student's strengths and weaknesses, and monitoring progress. Often known as formative assessment, this is a personal activity not designed to contribute either to a student's final evaluation or to the judgment of a school's performance. It is its individualised and confidential nature, and the fact that there are no external high stakes associated with it, that gives it its usefulness.

Assessment of a *summative* nature, is designed to *certify* an individual, to judge whether the individual is qualified to proceed to a further stage of education, training or employment. Public examinations are of this kind.

Assessment can be designed to provide information about the *performance of an educa-*

1070-3535/96/040199+ 05 $12.00/0

tional system, its institutions, or specific programmes and approaches. This use of assessment is essentially a research activity. It studies the nature of relationships between factors, by measuring student achievement in different circumstances, and by making allowance for background factors such as social status. It may be used to speculate about causal relationships, for example between class size and learning, or to evaluate the results of particular courses of action, such as the introduction of a new reading scheme.

These purposes are distinct. Attempts to utilise an assessment designed for one kind of activity for a different purpose may be inappropriate, invalid and possibly lead to unintended and undesirable consequences. One example is the use of examination results to judge the performance of schools by publishing rankings or league tables which is a distortion of the primary purpose of providing individual certification. It fails to provide reliable evidence about schools and also distorts the behaviour of schools in undesirable ways. Another example is in the compilation of written student records of achievement. In order to incorporate proper diagnostic information in these student records weaknesses as well as strengths would have to appear in the record: insofar as the records subsequently become used for judgmental or certification purpses, there will be an unwillingness to record weaknesses, other than euphemistically. One result is that objective recording of diagnostic information may tend to disappear, and the written record then becomes an inaccurate account, reflecting only positive achievements, whatever its benefits in terms of motivation and the like.

School league tables

The systematic publication of performance tables for public examination results, begun in 1992, is now an established feature of the educational system in England and Wales. At GCSE the most prominent feature is the presentation, for each school, of the percentage who achieve 5 or more passes at grades A-C: at A-level an average grade point score is produced for each school. The national and local press are encouraged to present school and college results ranked in terms of these percentages and averages, and the Parents' Charter encourages people to use these tables in choosing schools and colleges.

The principal argument against such a use is that the examination 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 such outputs. More recently the Government has accepted the inadequacy of using such crude rankings and that what is required are so called value added tables in which there is a proper allowance for pre-existing achievements (DFE 1995). Inconsistently, it continues to promote the use of the existing unadjusted tables.

In addition to this obvious inadequacy of existing performance tables there are other problems which apply also to value added tables. While these tables are an improvement and can be useful if read in conjunction with other information, some of the initial expectations that these could provide sensitive indicators of school performance have not been fulfilled.

The flaw of averages

One obvious problem with reporting only a single figure such as the overall percentage of high grades is that 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 vice versa for the other. Research on this topic has demonstrated that this does indeed occur (Sammons et al 1995). Likewise, where value added tables are concerned, some schools may exhibit relatively good

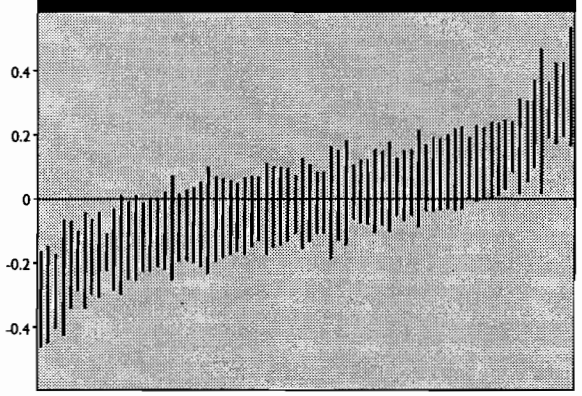
performance for initially poorly achieving students and produce relatively weak performance for initially highly achieving students and vice versa for another school. Such differences may be masked by the use of a single figure.

The problem of uncertainty

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 means that very little can be said about any individual departments 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.

Statistical uncertainty means it is hard to say which school does better

Range of statistical uncertainty for value-added at A-level by school



Source: Goldstein and Thomas 1996 Using examination results as indicators of school and college performance *Journal of the Royal Statistical Society*

The Chart 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. The lines represent ranges of statistical uncertainty such that it is possible to judge two schools or colleges as truly having different value added scores only when the lines do not overlap. In this Chart, 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 comparions are of limited value since in most cases we will find no difference.

Historical data

A further problem with all of these tables is that inevitably they refer to students who began their education at those institutions many years earlier. Thus, the GCSE results published in November 1996 refer to students starting at their secondary schools some five years previously: given that schools can change markedly over time there will be additional uncertainty over the use of those results to predict the performance of a cohort starting in 1997.

Social factors

Recent research (Goldstein and Sammons 1996) shows that the primary school attended by a child exerts an important influence on GCSE performance and that this should 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 performance tables, 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 ef-

forts at tracing them and recording their examination and test results.

Taking all these caveats together it is clear 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 to be further investigated, bearing in mind that the information is historical. These caveats refer not only to the public presentation of comparative tables but also to the use of such information for internal purposes by individual schools as is currently being proposed by the Schools Curriculum and Assessment Authority (SCAA) for Key Stage test results as well as public examinations. At the very least, if such comparisons are to be attempted, it is vital to provide users with careful descriptions of all their limitations. If this were done many users might heavily discount such evidence.

Negative feedback

Finally, the existence of league tables within a competitive marketplace has invested them with an extra importance. To have a high rank or to be improving is seen to be a competitive advantage and there will be pressure for schools and colleges to modify their behaviour to secure such an advantage. For example, a key statistic in reporting GCSE results is the percentage of subjects passed with grades A-C. By concentrating efforts on those students predicted to obtain GCSE subject grades around the C/D borderline a school may hope to increase the proportion of its grade A-C passes, but only to the detriment of relative neglect of the very low achieving or the very high achieving students. Whether intended or not, such distortions of education are an regrettable but inevitable consequence of such a high stakes accountability system.

As one source of information about school performance, league tables can have value, assuming that they are properly contextualised, at least by adjusting for intake achievement. They may indicate to LEAs for example where there are potential problems or examples of highly successful schools or departments which could usefully be followed up. They may be able to indicate, over time, where improvements or deteriorations are taking place and they can form a part of continuing research activities studying factors associated with performance. It would be unfortunate if such positive uses were to become obscured by the public promotion of league tables, value added or otherwise, as valid tools for judging schools and colleges.

International comparisons

Governments and international oganisations such as the OECD have begun to place considerable emphasis upon international studies (carried out largely by the International Association for the Study of Education Achievement (IEA)) as indicators of the quality of educational systems, and by implication of potential economic importance. A recent report based largely upon the IEAs second international maths and science studies of the early 1980s drew strong inferences about educational quality from comparing countries such as Taiwan and the UK and has been widely quoted (Reynolds and Farrell 1996). The study involved detailed work by experienced practitioners in educational assessment in devising tests and questionnaires for use in samples of schools in a range of countries. Forty countries were studied, necessitating extensive trailing of materials, careful attention to problems of translation and the co-ordination of sampling and administration. Information is collected from students (typically at ages of eight and thirteen years), teachers and schools and the students complete test forms. There is information on the content of the curriculum and how much has been covered. Some of the more interesting research is that which sets out to relate test performance, within narrowly defined areas, to curriculum exposure and to see

how this differs from country to country.

There are many difficulties in ensuring that the information obtained by such studies is satisfactory – such as those of adequate population coverage, sampling of schools and students and translation of tests and questionnaires. One of the principal ones is the fact that they are cross-sectional, that is they measure each student only once. The use of these data to make comparisons between schools, and likewise between countries is severely limited, since no account can be taken of the achievements the students had when they started the relevant phase of education, or indeed when they started formal schooling. Just as schools differ in terms of the intake attainments of their students, so whole educational systems may differ, for cultural, social, economic or political reasons. Without such intake information on the individual students, inferences about the effects of the educational systems *per se* will be unsafe.

There is a further major difficulty in ascribing causal relationships, and this derives from the fact that the countries involved differ along many dimensions – more than there are countries involved; for example in terms of employment prospects, cultural attitudes towards education, teaching styles, class sizes, teacher salaries etc. Thus, for example, to associate a difference between Taiwan and England in certain aspects of mathematics performance with, say, the organisation of teaching in whole class groups, while ignoring all the other ways in which those countries differ is a case of very special pleading and is scientifically invalid. While there is clearly a temptation for politicians and others to choose those interpretations compatible their pre-existing views, such inferences need to be resisted.

There is much of value to be learnt from these international comparative studies, and they should be encouraged. But they should not be used for the purpose of simple comparisons between countries – international league tables – or to bolster support for particular views about which factors promote learning. To do so risks discrediting the studies and diverting attention from the insights they do have to offer in describing how the processes of education differ among systems and in suggesting how future research might be focused.

Conclusions

In conclusion, we need to recognise the inherent complexities of making statements about the performances of teachers, institutions or whole educational systems. The only meaningful way to make judgments is with a full understanding of context and circumstances. Very little is to be gleaned from rankings and league tables. In the case of school league tables, on the one hand the use of assessments designed for certification of individuals to judge their school is totally inappropriate. On the other, the information obtained can be positively misleading. Tables of examination performnce condemn schools with low results even when they may be producing good performances from a very low-ability intake. At another level even a sophisticated value added analysis may be inadequate because there is too much uncertainty. Furthermore, institutions differ along many dimensions and attempts to rank them are likely to be gross oversimplifications.

To say that a task is complex and difficult, however, is not to say that it should be abandoned. There is much to be learnt from research into institutional and systematic differences but this effort risks being overlooked amid crude attempts to produce simple-minded measures for labelling institutions and countries. Those in authority, or who wish to be in authority, seem to have little taste for confronting the complexities of educational performance. Complaining about standards of education while at the same time perpetuating untenable or inappropriate claims is inconsistent and dangerous. The failure of politicians themselves to adopt rational standards of debate must surely undermine their purported attempts to promote those same standards in education ●