There are serious difficulties in the path of providing reliable rankings or league tables of schools

Assessment And Accountability

Harvey Goldstein
Institute of Education
University of London

THE assessment of students is now a major educational issue, whether this is done by teachers, end-of-course examinations or standardised tests. I shall describe how assessment has come to be used for accountability purposes, setting the scene with a summary of the four major roles of assessment.

1. Diagnostic or 'formative' assessment is the process whereby the knowledge and understanding of a student is assessed in order to evaluate his or her strengths and weaknesses so that both the student and the teacher can organise learning efficiently.

The assessment instrument is often administered individually, but it could be a class test: the key feature is that its purpose is to 'diagnose' the student's stage of learning.

It is a 'private' form of assessment because its results are not publicly available and are intended solely for the enhancement of the individual's learning. This distinguishes it clearly from other forms of assessment.

2. Asessment for selection purposes is the use of a test, teacher evaluation or public examination to grade students. The result of the grading is available for others to use in selection, for example for further stages of education or for employment. It differs in function from diagnostic assessment in that the aim of the students, at the time they are assessed, is to maximise their scores or grades, rather than expose their weaknesses for subsequent correction.

This means that the format of the assessment, the attitude of the student towards it and the way it is administered will all differ in important respects from diagnostic assessment.

Both of these types of assessment are concerned with what happens to the individual student. The last two types are concerned with making statements about groups of students.

3. Assessment for institutional accountability is the use of assessment results, from the students in a school (or other institution), in order to make statements about the performance of the school. We may use standardised tests, examinations or teacher assessments to attempt to do this. The results are 'public' in the sense that institutions may be held publicly accountable for them.

4. The fourth purpose of assessment is to make comparisons at a system level — for example, between different kinds of schools or between, say, different methods for teacher reading. This purpose may be characterised as a research function, since it is trying to draw general lessons without naming individual pupils, institutions or authorities.

In the UK its most important recent manifestation was in the government-sponsored but now defunct 'Assessment of Performance Unit'.

If an assessment system is to prosper and if it is to retain intellectual integrity, it must avoid claiming that it can serve conflicting aims simultaneously

Characterising assessment in terms of its function provides a key to understanding much of the current debate. The actual form of an assessment is of secondary importance and for the most part is a somewhat technical matter. In principle, any form of assessment can be used for any of the above purposes. It does not matter, for example, whether the assessment is labelled as 'criterion' rather than 'norm' referenced. What counts is how it is to be used. The purpose determines how students react and how teachers and schools respond in terms of curriculum content, form and organisation.

Finally, it is difficult to see how the various purposes can be accomodated within a single assessment. If an assessment is to be formative, students must be encouraged to expose their weaknesses, and this is not encouraged by a selection examination; nor is it encouraged by assessment which is designed to hold schools accountable for the performance of their students.

If an assessment system is to prosper and if it is to retain intellectual integrity, it must avoid claiming that it can serve conflicting aims simultaneously.

An important reason for the failure of National Curriculum Assessment in the early 1990s was just this advocacy of assessments serving several conflicting aims. The responsibility for this can be traced back to the Task Group on Assessment and Testing (TGAT) which reported in 1988.

The use of exam or test scores to compare schools or education authorities is practised not only in the UK but in many other educational systems. At first glance, this use of assessment for public accountability has some validity. After all, a major purpose of schools is to enhance the understanding and knowledge of their students, so that it seems quite reasonable to judge schools on this basis.

Of course, we can debate the appropriateness of different kinds of assessment, whether carried out by teachers or, for example, by rigorously applied standardised tests. We can also have a useful debate about exactly what is to be assessed for this purpose. In what follows I shall concentrate on the general validity of any assessment designed for accountability purposes.

The first, perhaps obvious point to be made is that the knowledge and understanding that students attain is determined not only by what they experience at school, but also by their experiences before they entered any particular school, in their homes, in society at large, and as a result of their biological or genetic make-up.

Thus, if we compare two secondary schools whose intakes differ considerably in terms of the pre-existing performance of their students, we would expect the students in the school with the higher-performing intake to perform better on any subsequent assessment. We expect this because we realise that the other school would need to work harder to achieve the same result.

On the other hand, if the students at each school were very similar in

It is important to avoid the trap of supposing that the provision of some information about schools, even just unadjusted raw league tables, is bettter than no information at all

their achievements on entry, then we would generally find it more acceptable to judge the schools in terms of the later performances of their students. It is this reasoning that lies behind attempts to carry out so-called 'value added' comparisons.

In simple, non-technical terms, value-added analysis works in the following way. Suppose we have two schools: school A that has a high-scoring intake and school B with a low-scoring intake.

Although the schools differ in overall intake performance, there will not normally be a sufficiently wide range of intake scores in each school so that there are some students at each point of the score scale. This means that we can compare the subsequent assessment results for each group of students who have the same intake score, and we can do this for the whole range of scores.

Thus we can see whether the low-scoring students in school A subsequently do better or worse than the low-scoring students in school B, and so on. The point is that by allowing for or adjusting the initial scores in this way we will be comparing like with like.

This is really all there is to value-added analysis. Sometimes we will find that, for all the initial intake scores, school B students perform consistently better than those from school A. In such cases we might average the differences between the scores from each school over the whole range of intake scores and quote a single 'value-added' comparison.

In other cases we may find that, say, for low intake scores the students in school A do better on average than those in school B but vice-versa for high intake scores. Such differential value-added performance is both interesting and important when it occurs.

Some commentators have suggested that value-added comparisons can be carried out simply by forming the arithmetic difference between an 'outcome' assessment score and an intake score. This has been raised in connection with the existing 10-point scale for National Curriculum Assessment where a simple difference between levels achieved has been proposed.

One reason why such a procedure is unviable is because it does not allow for the possibility of differential value-added performance.

Whether we have carried out a value-added analysis or even a simple unadjusted analysis, all comparisons between schools will be subject to a margin of statistical 'error' or 'uncertainty'. This will be relatively large when the number of students in a school is small.

For example, the Newcastle ALIS scheme that carries out value-added analyses of A-level results reports comparisons on a subject-department basis where numbers may be very small indeed.

Work at the London Institute of Education has shown that for the GCSE examination, value-added analyses yield 'uncertainty intervals' which are so large that only schools with extremely small or extremely large value-added scores can be separated from the remainder. In other words, it is not technically possible with any reasonable certainty to give an unequivocal ranking of schools in a league table, whether this is an unadjusted 'raw' table or an adjusted value-added one.

Finally, besides the need to take account of intake assessment, it may also be important to take account of other factors, outside the general control of the school, which may be responsible for influencing performance. Home environment is clearly one of these, but so are other factors such as resources available for teaching etc.

If there is a political wish to find ways of comparing how schools set out about their tasks, then there are other perfectly proper and direct methods for so doing

I hope that this brief outline of some key issues has made it clear that there are serious difficulties in the path of providing reliable rankings or league tables of schools. Not only are comparisons that are based upon raw results misleading and potentially unfair, we should not expect even value-added analyses to provide definitive comparisons. Furthermore, value-added analyses, although they are certainly an improvement, are necessarily more complicated to carry out. They require intake assessments, and records for individual students have to be linked together over their school careers — an expensive and time-consuming activity.

There is one further, fundamental problem with using assessments to make judgements about schools: namely, that such judgements can be seriously out of date. For example, if we were to use GCSE results for the 1994 cohort of students (whether value-added or not) to judge a school's performance, this would refer to a group of students who had entered their schools five years previously in 1989. Because schools can change markedly over such a period, the use of 1994 results to predict the performance of a cohort who enter school in 1994 and who therefore will take exams in 1999, is likely to be a perilous business.

I am not against value-added analysis, I think it is important to carry out such analyses for research purposes in order to study the factors that are responsible for some schools appearing more effective than others for particular groups of students. My doubts are whether the results of such analyses can be used to hold schools accountable.

EVERYTHING I have said about the use of assessment results for accountability purposes applies with equal force, and for similar reasons, to other measures such as truancy rates or student attitudes. If there is a political wish to find ways of comparing how schools set out about their tasks, then there are other perfectly proper and direct methods for so doing. Such methods range from the use of highly-trained inspectors to systems of teacher self-assessment.

Finally, I believe it is important to avoid the trap of supposing that the provision of some information about schools, even just unadjusted raw league tables, is bettter than no information at all.

The problem is that such information will be biased if it does not adjust for intake and it will convey a spurious accuracy if it is presented without the corresponding 'uncertainty intervals'. It may be more informative to say nothing, rather than run the risk of misleading people.

At the very least, if such information does get published, we should insist upon the normal canons of scientific evidence and social integrity. These require that stringent warnings with clear explanations are displayed prominently, reminding the reader that the numbers should in no way be construed as reflecting well or badly upon the institutions involved. PB