British Educational Research Journal, Vol 27, No. 4, 2001

Using Pupil Performance Data for Judging Schools and Teachers: scope and limitations

HARVEY GOLDSTEIN, *Institute of Education University of London*

ABSTRACT *The article reviews ways in which performance data are currently used within the England and Wales education system. It uses research evidence to critique many of these uses and presents some alternative proposals for a more rational approach.*

Sources of Performance Data

Performance information for school age children traditionally has been available using public examination results at the end of compulsory schooling, and also at A level. These were also the first data to be used as a source of school comparisons, starting in 1992. Since the mid-1990s, Key Stage test results have also become available and are now published annually for all schools. A description of these data and the Government's recommendations for their use can be found in the 'Autumn Package' issued each year (Department for Education and Employment (DfEE), 1999).

All of these data are available for all schools either as a by-product of the public examining system or as a result of central government initiatives to provide 'accountability' measures. In addition, there are several local education authority (LEA) based schemes of assessment that are used within local areas; one such scheme will be described in more detail later.

Surveys and research studies also collect assessment data. These will typically have associated with them other measurements on pupils, their teachers, their schools and backgrounds, but the major use of such studies is for research into the understanding of schooling. Because they are collected on a sample basis, they are also unsuitable for use as a means of comparing institutions.

The notion of 'accountability through pupil assessment' became prominent in the 1980s with the setting up and development of the Government's Assessment of Performance Unit (APU) (see Gipps & Goldstein, 1983). By the end of the 1980s, the APU had been dismantled and 'accountability' came to be associated with the performance of individual schools rather than the system as a whole. Whereas the APU was

Received 26 September 2000; resubmitted 16 March 2001; accepted 20 March 2001.

ISSN 0141-1926 (print)/ISSN 1469-3518 (online)/01/040433-10 ©2001 British Educational Research Association
DOI: 10.1080/0141192012007144 3

concerned with issues such as differences between boys and girls, changes in performance over time and the effects of styles and formats of test instruments, the new accountability narrowed its focus to simple comparisons of schools in terms of average scores in a small number of 'key skills'. The view of the then Conservative administration, as well as the subsequent Labour administration, was that such comparisons were adequate for the purpose of providing a yardstick by which parental choice of schools could be assisted. More recently, these comparisons have become associated with achievement 'targets' and ultimately as a means of allocating resources towards individual schools and teachers via appraisal and promotion procedures.

An important foundation for these accountability measures in the eyes of government is the notion that assessments based upon centrally controlled tests are 'objective' and also reliable enough to provide acceptable criteria for comparisons. Both these assumptions are questionable. The reliability issue has been discussed widely in the context of the uncertainty associated with 'league tables', and I will return to it later. The claim for 'objectivity', however, has less often been challenged and is worth further study.

A typical claim by the proponents of 'standardised' or 'centralised' tests in education is that their separation from the sources of instruction lend them an 'objectivity' that, for example, teacher constructed tests do not have. Some proponents of tests based upon 'item response models' go further and refer to what they do explicitly as 'objective measurement' (see, for example, Lord [1980] for a detailed exposition). What is intended by the term 'objective' is to place such tests on the same basis as commonly used physical measuring instruments, such as a stadiometer for measuring height, which are accepted as providing valid and interpretable comparisons among individuals, and in particular to avoid the introduction of personal characteristics of the measurer into the process. Nevertheless, the difficulty of applying this notion to tests of educational achievement is essentially twofold.

First, since individual learning environments differ and pupils will understand different aspects of a subject, the actual content of any test (or series of tests) is relevant. Thus, two pupils may have learnt different aspects of a period of history, but if the content of a test does not reflect what one of them has learnt, then the test could be said to be biased against that pupil. A common response to this issue is in fact to require a degree of uniformity in terms of what is taught, and this is the basis, for example, of public examinations and increasingly of the National Curriculum in England and Wales. Even so, as Cresswell (2000) points out, a large element of 'subjectivity' remains within such systems. The second problem is that there is now extensive research (Foxman *et al.*, 1990) that demonstrates how apparently innocuous features of a test, such as the physical layout or ordering of questions, can affect responses, sometimes markedly. It would seem that those who are tested are not merely passive objects against whom a measuring instrument can be placed, but interact with the characteristics of the instrument in a way that undermines the usual claims for neutrality. For both these reasons a great deal of responsibility for the outcome of a testing programme relies upon the beliefs and views of those who construct the tests, and here again we can expect variability which makes claims to 'objectivity' questionable.

These problems become acute when there are attempts to use test results to make comparisons over time. In public examinations and Key Stage tests, different instruments are used over time, so that in order to make statements about changes in performance over time, it is essential that a common scale can be constructed for the various tests that are used. The difficulty is that, because of a lack of a common 'objective' yardstick, there is no unique way to do this. Whichever way it is done, it is not possible to decide

whether any change in test score is *really* due to a change in performance or a change in the difficulty of the test, or a mixture of the two (see Goldstein [2000] for a more detailed discussion of this issue). Thus, when policy-makers base policy and 'rewards' upon changes in 'standards' (test scores), they have no way of knowing whether what is observed is real or not.

In the following sections, I will be looking at the particular problems posed by the use of performance data for judging schools and teachers, but it should be remembered that the 'objectivity' issue remains even when some of the other problems associated with making judgements are resolved.

Performance Data Used for School Comparisons

In England and Wales a pattern has now been established whereby, every autumn, the test and examination results for every school at Key Stages 2, 4 and A level are published and national results at Key Stages 1–4 and A level. These form the basis for 'league tables', rankings of schools published nationally and locally by the media. There is also now a recognition of the severe limitations of such 'raw' comparisons, which fail to take into account the prior intake achievements of pupils, and a recent innovation is the provision of limited 'value added' data from Key Stage 3 to Key Stage 4, although it is not entirely clear whether this publication will be continued. I shall go into more detail about value added comparisons later, but although this innovation is welcome, it is less than adequate on a number of grounds.

Firstly, it does not properly adjust for intake since pupils have already been in secondary school for over 2 years by the time Key Stage 3 tests are taken. Secondly, while, in principle, a school can estimate an uncertainty (confidence) interval for its value added score, there are few details in the Autumn Package (DfEE, 1999) about how this might be done. While schools are encouraged to study and plot their pupils results for Key Stage 4 against those for Key Stage 3, it is difficult, using a single school's data, to detect where there are large departures from the overall pattern: this can most efficiently be done by carrying out value added analyses on schools as a whole (see later).

In its current state, the Government's advice to schools on using performance data is at best confusing and at worst misleading. Thus, for example, for a school's raw scores in comparison to the national average, schools are encouraged to 'identify any features of subject organisation or teaching practices that they feel particularly contribute to their successful results', and to 'note features where 'pupils' achievements are below par'. Clearly, the assumption here is that simple, unadjusted average comparisons can identify 'successful results', yet elsewhere, the Autumn Package extols the use of value added adjusted scores as being a sound basis for comparisons; it is not possible to have it both ways.

The Autumn Package also presents 'benchmark' information, by which is meant summary data within five categories defined by the percentage of pupils eligible for free school meals. The idea is that schools will be able to make comparisons against results from schools whose pupils come from 'similar' backgrounds. Again, as I will illustrate, the research evidence shows clearly that while such comparisons are more valid than unadjusted ones, they are far from adequate when compared to full value added comparisons. Official government publications make little attempt to place such analyses in context, and give no hint that this is problematical.

While it is true that, in some of its public statements, the DfEE has stated the

desirability of moving to a fairer system of comparing schools using value added data, its failure to set out the known limitations of its current offerings does not inspire confidence that future offerings will provide a more honest presentation.

Performance Data Used in Office for Standards in Education (OFSTED) Inspections

When OFSTED inspection teams prepare for inspections, they are given a great deal of information about the performance of pupils on Key Stage tests and examinations. For primary schools, for example, they will have available the average scores in mathematics, English and science for the most recent and preceding cohorts at KS1 and KS2. Inspectors are expected to use such information, together with other 'contextual' factors such as the proportion of pupils eligible for free meals, to make preliminary judgements about the 'performance' of the school. This information is provided in so-called performance and assessment (PANDA) reports, which are available to the school. Inspection teams are expected to use this information as part of the process of judging the performance of the school and its teachers. I have already pointed out the inadequacy of such information, so that, as with schools, it is not clear just how such limited information can be used constructively while recognising its defects.

Until 1999, the guidance to inspection teams stressed the primacy of performance data and in particular that inspectors would need to provide strong justification for judgements that appeared to dispute the performance of the school as represented by the unadjusted test results. There are some signs, however, that this guidance is changing and that a more realistic view is being taken (Time Educational Supplement, 2000), but this will require a proper training of inspection teams in the interpretation of performance data and a familiarity with the issues surrounding value added analysis and uncertainty.

Performance Data Used in Teacher Appraisal

In September 2000, new arrangements for monitoring the performance of teachers were due to come into force (*www.dfee.gov.uk/teachingreforms/*). These have aroused considerable controversy and opposition from teachers' representatives (see, for example, the National Union of Teachers' website—*www.teachers.org.uk/*). The various documents describe how teachers are meant to use measures of pupil progress to support annual judgements of their performance—the *performance management framework*, or for an application to meet the *threshold standards* for promotion to a higher pay scale. Because of a successful legal challenge, the implementation of these has been delayed and it is now not clear what will emerge, although the Government still intends to introduce these arrangements.

The government documents place a considerable emphasis upon pupil progress as a means of judging the performance of teachers. They explicitly link teachers' achievements of progress targets with the targets that have already been set at school level, for example, in terms of percentages of students reaching Key Stage test levels at various ages. There appears to be some confusion in this respect, since while the *Performance Management* documents state that it really is *progress* that is to be assessed, taking into account pupils' prior achievement, the school level targets generally have not been set on that basis. This confusion is also evident in the documents themselves; in examples of performance management objectives (see the DfEE 0051/2000, Performance manage-

ment framework) in the five examples given, reference is made to pupil progress in each case, but in only one of these is there even a hint that prior achievement is to be taken into account—the others refer to judging the 'absolute' achievements of a class or group. On the other hand, the documents on *Threshold Assessment* are clearer and do explicitly refer to 'value added' assessment, although they too betray a lack of understanding about what such assessments can be used for.

The Annual Performance Review

The documentation on judging teacher performance (Annex C on setting objectives) says that 'what is important is that the planning discussions are based on an understanding of pupils' prior attainment'. Nowhere, however, is there a discussion of just how this is to be done, or its limitations. In particular, for any one teacher, the prior attainment most relevant to pupil achievement at the end of a school year is that at the start of the same year. The problem is that there are few data available for measuring progress over a single year—most of the value added studies are over longer periods between Key Stage testing times. Without LEA or national norms, it is difficult to see how any sensible judgements can be made—leaving aside the problem of the wide uncertainty intervals that might be expected for such yearly comparisons. If sound value added judgements cannot be made, then presumably teachers will be assessed on 'absolute' achievement levels—and it is therefore perhaps no surprise that the examples given are mainly of this nature.

Throughout these documents there is the implicit assumption that teachers and schools alone are what influence the achievements of their pupils. Yet, not only is prior attainment important, so is the education at any previous schools attended, mobility, special needs etc. Moreover, especially in secondary schools, it is very difficult to ascribe the progress of any one pupil in a given subject to the teacher of that subject. Pupil progress will be affected by other teachers and features of schooling and background—perhaps to an even greater extent—and the attempt to associate progress for a pupil with a single teacher is not only divisive, it is also likely to be misleading.

Threshold Applications

In applying for threshold promotion, teachers 'should show that as a result of their teaching their pupils achieve well in relation to their prior attainment, making progress as good or better than similar pupils nationally'. Teachers are expected to use national test data 'where appropriate'. When this is not done, and as pointed out earlier this will commonly be the case, it is stated that alternative valid methods for showing progress are 'examination marks, general coursework marks etc.' Quite how these are supposed to show progress, when they generally will not take into account prior achievement, is not stated.

In discussing the evaluation of progress, several sources of evidence are referred to. One of these is the PANDA reports, in particular where these attempt to make adjustments on the basis of free school meal entitlement, and I shall have more to say later about this kind of adjustment. Another reference is to the various value added schemes, such as A Level Information System (ALIS) (FitzGibbon, 1992), Performance Indicators in Primary Schools (PIPS) Tymms *et al.*, 1997) and Kendall (1995), which make comparisons between schools and departments after adjusting for prior achievement. As already pointed out, these will typically not refer to the relevant period, a school year, and are only available for a minority of schools, and it is difficult

to see how they could be used to make a judgement about any individual teacher. There is reference to Qualifications and Curriculum Authority (QCA) optional tests that produce scores every year, at least for primary children, that could be used. These potentially would be the most useful information, but there is still no recognition of the limitations of such information. The documents refer frequently to the 'expected acceptable range' of progress but nowhere define this. The uncertainty surrounding any measures, especially when based upon one class for one teacher, is very large and a proper contextualisation of any measure would need to involve all the other influences that a child is subject to, from the school as a whole, from other teachers and from external factors.

In short, in terms of pupil progress, it seems that teachers are being asked to provide evidence that, in many if not most cases, they simply will be unable to do in any objective fashion. In particular, a teacher may indeed have provided conditions in which pupils have made outstanding progress, but through no fault of their own be quite unable to demonstrate that fact.

Adjusting Performance to Reflect Progress—value added data

In addition to the schemes referred to above which have been important pioneers in providing value added information to schools, Hampshire LEA has been collecting and analysing primary school KS1 and KS2 test results since the early 1990s. It has been carrying out value added analyses that are fed back to schools as additional information for them to judge their performance, and details of this are discussed in the next section. More recently, it conducted an analysis for OFSTED which looked at the varying judgements about schools that would be made using 'raw' test results, benchmark data using free school meals and a full value added analysis (available from *www.ioe.ac.uk/hgpersonal/*).

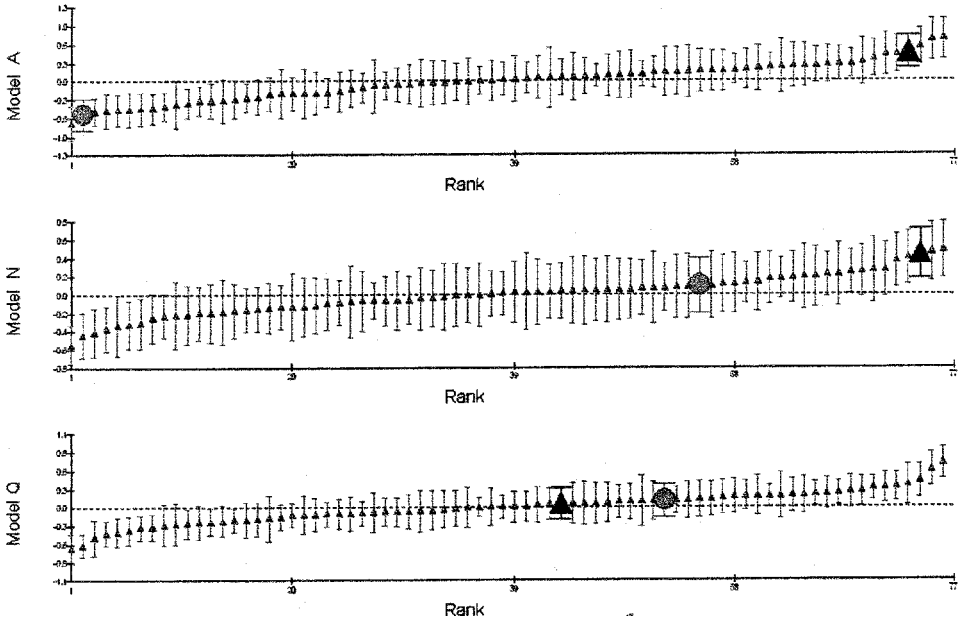

FIG. 1. Mathematics at KS2 (1998): value added (residual) plots with 90% confidence intervals.

Fig 1 demonstrates how very different inferences can be made depending on the procedure used. It is intended only to illustrate that it is possible for individual schools to change position, rather than to suggest that this is the norm. In this figure, Model A is simply the raw KS2 Mathematics test scores for schools ranked, together with 90% confidence intervals. Model N uses in addition an adjustment for free school meals, using information from both individual pupils and for the school as a whole. Model Q is a value added analysis which, in addition to the free school meals variables, uses KS1 test scores to adjust KS2 scores. The relative positions of two schools are shown, represented by a circle and a triangle. The triangle school is one of the 'best performing' in terms of raw score, with a confidence interval that does not cross the (dotted) mean line. It remains so when free school meals are taken into account, but in the value added analysis it is near the centre of the distribution and not distinguishable from the mean. The circle school, on the other hand, is significantly below average in terms of its raw score, indistinguishable from the mean when free school meals are allowed for and likewise for the value added analysis, where it actually comes out slightly better than the triangle school.

Similar results are found for English and it is clear that for any particular school both raw scores and 'benchmarked' scores can be misleading. It should also be observed that the confidence intervals in most cases include the overall mean so that it is anyway only at the upper and lower ends that schools legitimately can be statistically differentiated from the average. Furthermore, it may well be the case that the statistical 'models' that produce value added scores may lack information on further important variables, such as student background characteristics. Also, Goldstein & Sammons (1997) showed that for General Certificate of Secondary Education (GCSE) results at secondary level, the junior school attended was also an important predictor of performance. This raises the possibility that conventional value added analyses which adjust only for achievement at entry to the current school may be inadequate. In addition, schools will generally exhibit differential value added scores, where performance for, say, low achievers, is different to that for high achievers. This is elaborated in the next section.

Finally, a pervasive problem for all analyses of performance is pupil mobility. Pupils who change schools tend to have lower test scores on average. If a school has a high proportion of such children, then raw scores will tend to lower perceived performance and value added scores may have an upward bias if computed solely from those pupils present at the start and end of the particular stage of schooling. This is one more issue that requires the use of caution when interpreting results.

Private versus Public Accountability through Performance Data

The Hampshire value added scheme has already been mentioned (see Yang *et al.* [1999] for details). The aim of this is to produce information for schools that captures as much as possible of the progress made by individual pupils, adjusting for various characteristics and prior achievements. At the same time, there is a concern that the data should not be overinterpreted and ways have been devised to present it in accordance with this.

A key feature of the scheme, which arose from requests by the schools themselves, is that the results remain private. Each school sees its own results in comparison with those for other schools but these others are not individually identified. The results are not published for use as 'value added league tables' and so do not form part of a public accountability system. The view is taken that published league tables lead ultimately to attempts to 'play the system' whereby every effort is made to improve a school's

position, often in conflict with desirable educational practice. Such effects would be expected to occur whether raw *or* value added results were being used. The upshot would be results that would, at least partly, reflect the ability of schools to manipulate the system, rather than just reflecting the quality of education. The value added data in Hampshire, therefore, are viewed as additional information that schools, with assistance from the LEA, can use for improvement or change purposes. As such, it runs counter to current government policy that implicitly requires public comparisons.

The data used in Hampshire, in addition to prior achievements (KS1 for KS2 results and baseline measures for KS1 results) includes information on gender, free school meal entitlement, formal special educational needs ascertainment, changes of school, and average achievements of all the pupils in the school at the time of entry. An example of information that is given to the school following a value added analysis carried out by the LEA is shown in Fig. 2.

Two features are notable. First, the school line is not parallel to that for the whole county. This 'differential effectiveness' is a common feature of value added analyses and means that schools may achieve quite different results for the initially low as compared to the initially high achievers. Secondly, the confidence interval is wide and in particular does not overlap the county line for those pupils initially below the mean but does do so for those above, so that it is only for the relatively low initial achievers that the mathematics results present evidence of a poorer than average performance.

Schools are encouraged to examine such plots, one in each subject, to see whether an extreme result is explicable, for example, in terms of specific events such as teacher illness, or whether there may be further factors which need investigation. It is intended that the non-public nature of the information will encourage honest attempts to understand and improve. Currently, Hampshire is looking at ways of presenting data over long time periods so that schools will be able to view trends. Further information about Hampshire, including other initiatives, is available from the DfEE website (*http://www.standards.dfee.gov.uk/lea/hampshirecouncil1 -1.html).*

Further Developments and Limitations

In this section, I want to look at a couple of further developments that could change the ways in which performance data are used.

Unique Pupil Numbers

In 1999, the DfEE introduced a scheme for assigning a single, 13 character identifier to each pupil in compulsory education which will allow a centralised database to be created which would track pupil movements within the system, as well as test and examination results (www.dfee.gov.uk/circulars/dfeepub/oct99/131099/index.htm). Eventually, even though the amount of personal information would be limited, such a database would allow detailed value added analyses, taking into account mobility and achievement at previous phases of education. The success of such analyses will depend on the coverage and quality of the database and this remains to be ascertained.

Annual Testing

(www.qca.org.uk/cgi-bin/news.pl?Range=All&Format=Standard&Terms=optional+test)
QCA has instituted a series of tests in reading, writing and mathematics to be taken

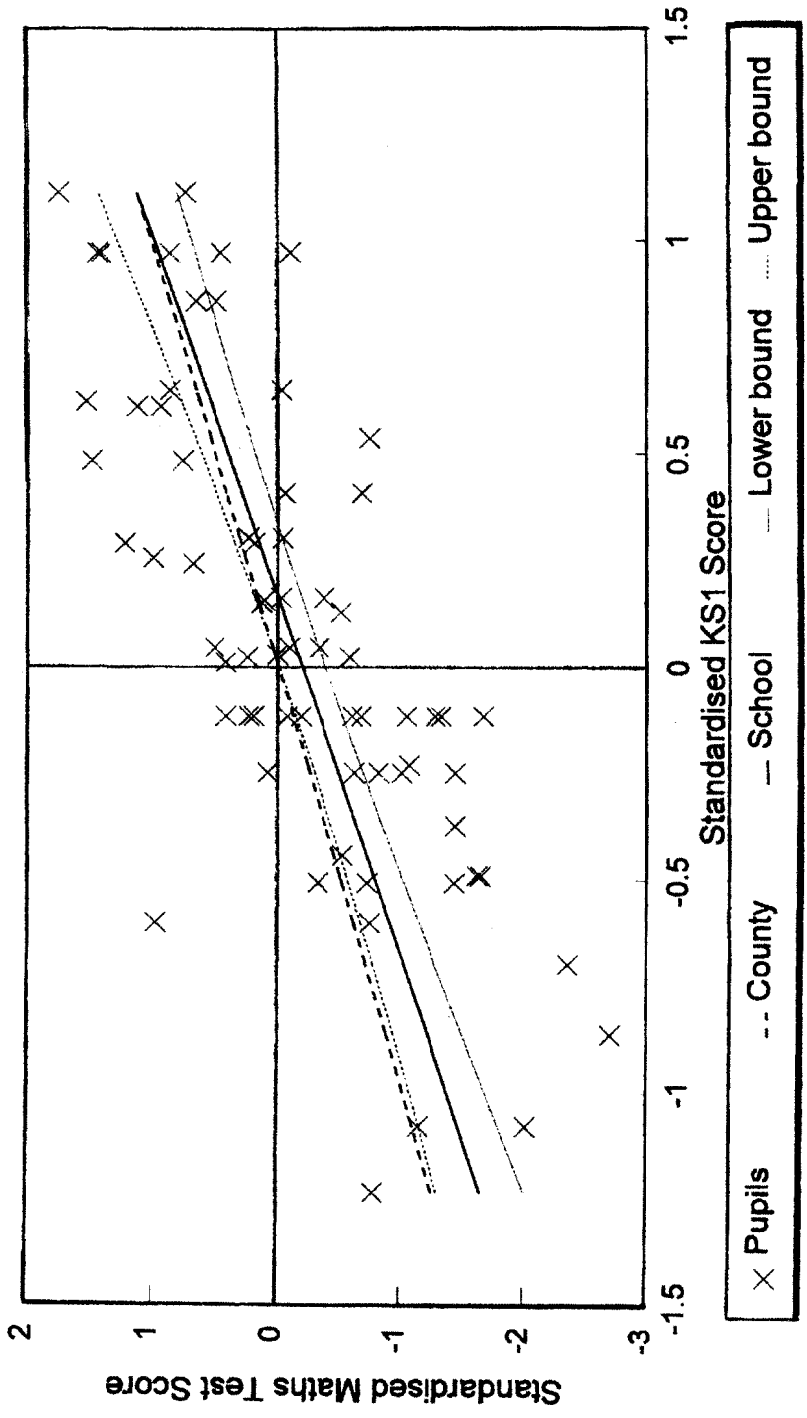

FIG. 2. A differential value added graph for one school at KS2. The County (LEA) line is drawn to pass through the origin and each school's line is drawn in relation to this using the residual estimates. For individual pupils the raw or total residual from the fixed part of the model is used and added to the County line at the appropriate KS1 composite score for that student.

at the end of years 3, 4 and 5 to complement Key Stage 1 and 2 results taken at the end of years 2 and 6. Using a sample which is stated to be nationally representative, it has produced norms for judging individual progress over one or more years. There are still technical problems associated with standardisation of the teacher scoring and the analysis and presentation are not particularly efficient, but there does seem to be some potential for more extensive and sophisticated (multilevel) analyses that would allow judgements associated with individual classes. An advantage of the (currently) optional nature of the tests is that they do not form part of an accountability system, and so, like the Hampshire scheme, could be used in a non-threatening fashion for purposes of school change.

Conclusions

It will be clear from this discussion that there is a great deal of inappropriate use of performance data for judging schools and teachers. Despite much research dating back to the 1980s, government has largely ignored its findings about the limitations of such data in terms of its practice, even while accepting the limitations in theory. Even so, there has been some recognition, within bodies such as QCA and OFSTED, as well as among educationalists at the DfEE, that a higher degree of sophistication is required, that value added data should be encouraged and schemes set in place that would collect useful data and allow more sensitive analyses. Schemes by LEAs, such as Hampshire, Lancashire and Surrey, and the ALIS projects provide models for future directions.

Nevertheless, it is difficult to see that any of the more useful schemes can really thrive while teacher appraisal based upon pupil achievements, school targets and annual league tables persists. What is required is a commitment to phasing out current procedures which serve a purpose which is largely politically driven, which is widely viewed as irrelevant and which, in its misleading nature, may be doing fundamental harm to education.

Correspondence: Harvey Goldstein, Institute of Education, University of London, UK; e-mail: h.Goldstein@ioe.ac.uk

REFERENCES

CRESSWELL, M. (2000) The role of public examinations in defining and monitoring standards, in: H. GOLDSTEIN & A. HEATH (Eds) *Educational Standards* (Oxford, Oxford University Press).

DEPARTMENT FOR EDUCATION AND EMPLOYMENT (DFEE) (1999) *The Autumn Package* London, DfEE (see for example, *www.standards.dfee.gov.uk/performance/html/KS2 contents.html*).

FITZGIBBON, C. T. (1992) School effects at A level: genesis of an information system? in: D. REYNOLDS & P. CUTTANCE (Eds) *School Effectiveness, Research Policy and Practice* (London, Cassell).

FOXMAN, D., RUDDOCK, G. & MCCALLUM, I. (1990) *APU Mathematics Monitoring 1984–88 (Phase 2)* (*London, Schools Examination and Assessment Council*).

GIPPS, C. & GOLDSTEIN, H. (1983) *Monitoring Children* (London, Heinemann).

GOLDSTEIN, H. (2000) Discussion of 'The measurement of Standards', in: H. GOLDSTEIN & A. HEATH (Eds) *Educational Standards* (Oxford, Oxford University Press).

GOLDSTEIN, H. & SAMMONS, P. (1997) The influence of secondary and junior schools on sixteen year examination performance: a cross-classified multilevel analysis, *School Effectiveness and School Improvement*, pp. 219–230.

KENDALL, L. (1995) *Examination Results in Context 1994: the NFER/LEA/AMA Project* (Slough, National Foundation for Educational Research).

LORD, F. M. (1980) *Applications of Item Response Theory to Practical Testing Problems* (Hillsdale, NJ, Lawrence Erlbaum Associates).

Times Educational Supplement (2000) Flawed OFSTED measure attacked, 14 July.

TYMMS, P., MERRELL, C. & HENDERSON, B. (1997) The first year at school: a quantitative investigation of the attainment and progress of pupils, *Educational Research and Evaluation*, 3, pp. 101–118.