Education Policy Analysis Archives

Volume 10 Number 6 **January 16, 2002** **ISSN 1068-2341**

A peer-reviewed scholarly journal
Editor: Gene V Glass
College of Education
Arizona State University

Articles appearing in **EPAA** are abstracted in the *Current Index to Journals in Education* by the ERIC Clearinghouse on Assessment and Evaluation and are permanently archived in *Resources in Education*.

Technical and Ethical Issues in Indicator Systems: Doing Things Right and Doing Wrong Things

Carol Taylor Fitz-Gibbon
University of Durham

Peter Tymms
University of Durham

Citation: Fitz-Gibbon, C.T. & Tymms, P. (2002, January 16).Technical and ethical issues in indicator systems: Doing things right and doing wrong things. *Education Policy Analysis Archives*, *10*(6). Retrieved [date] from http://epaa.asu.edu/epaa/v10n6/.

Abstract

Most indicator systems are top-down, published, management systems, addressing primarily the issue of public accountability. In contrast we describe here a university-based suite of "grass-roots," research-oriented indicator systems that are now subscribed to, voluntarily, by about 1 in 3 secondary schools and over 4,000 primary schools in England. The systems are also being used by groups in New Zealand, Australia and Hong Kong, and with international schools in 30 countries. These systems would not have grown had they not been cost-effective for schools. This demanded the technical excellence that makes possible the provision of one hundred percent accurate data in a very timely fashion.

> An infrastructure of powerful hardware and ever-improving software is needed, along with extensive programming to provide carefully chosen graphical and tabular presentations of data, giving at-a-glance comparative information. Highly skilled staff, always learning new techniques, have been essential, especially as we move into computer-based data collection. It has been important to adopt transparent, readily understood methods of data analysis where we are satisfied that these are accurate, and to model the processes that produce the data. This can mean, for example, modelling separate regression lines for 85 different examination syllabuses for one age group, because any aggregation can be shown to represent unfair comparisons. Ethical issues are surprisingly often lurking in technical decisions. For example, reporting outcomes from a continuous measure in terms of the percent of students who surpassed a certain level, produces unethical behavior: a concentration of teaching on borderline students. Distortion of behavior and data corruption are ever-present concerns in indicator systems. The systems we describe would have probably failed to thrive had they not addressed schools' on-going concerns about education. Moreover, data interpretation can only be completed in the schools, by those who know all the factors involved. Thus the commitment to working closely and collaboratively with schools in "distributed research" is important, along with "measuring what matters"... not only achievement. In particular the too-facile interpretation of correlation as causation that characterized much school effectiveness research had to be avoided and the need for experimentation promoted and demonstrated. Reasons for the exceptionally warm welcome from the teaching profession may include both threats (such as the unvalidated inspection regime run by the Office for Standards in Education) and opportunities (such as site based management).

Indicator systems that we have developed over the last 15 years have, somewhat to our surprise, attracted support and subscriptions from about a third of the schools in England where we work on a scale many times greater than any other group.[1] We have also developed a linked Curriculum, Evaluation and Management Centre at the University of Canterbury in Christchurch, New Zealand, and we have the pleasure of welcoming participation from scattered schools in thirty countries. The development of the indicator systems in the CEM Centre is unusual if not unique in that schools themselves have chosen to participate. The systems are therefore professional, ground-up, developments, that stand in contrast to "top-down" indicator systems created, and sometimes imposed, by state and local authorities. The interests of both kinds of systems should ultimately, however, be coincident: to improve education.

An indicator can be defined as an item of information collected at regular intervals to track the performance of a system (Fitz-Gibbon, 1990). The indicator systems that have formed the basis of our learning are all designed *to feed back valuable information* of interest to teachers and administrators in schools and colleges. We see our indicator or information systems as significantly empowering schools as they participate with a university in 'distributed research'. The issue of public accountability must also addressed by indicator systems.

The design of each system has been driven by the need to measure outcomes that matter along with relevant covariates so that fair comparisons can be made. Process variables are measured in some of the systems, but only to generate hypotheses, not to make judgements.

'Value added' measures have been included in our systems since the first one started in 1983. In 1995 we won the two year contract to conduct national feasibility studies for a value-added system. The final report, Fitz-Gibbon, 1997, can be found on the website for the Qualifications and Curriculum Authority (http://www.cem.dur.ac.uk/ca/5-14/durham_report.asp—click publications and search for "national" or under Software on the CEM Centre website). The studies carried out contributed to ongoing debates of both a technical and ethical nature. The issues are of interest to those concerned with indicators in any system, either as designers, users or policy makers.

Technical Issues

Technical issues include those procedural problems that must be solved if an indicator system is to be of high quality and run in a timely, efficient and effective fashion.

Indicators need to be based on adequate samples, have appropriate levels of reliability, good validity, and above all, positive reactivity. These are technical terms straight out of research methods courses, but they go to the heart of indicator systems, and any practical use of data. The data must be of research quality, otherwise it will confuse rather than guide.

Technical infrastructure

In the early years, in 1983, technical sophistication was no more than batch-processing using a mainframe. The mainframe could deliver capital letters and stars and eventually an adequate type-face could be constructed via embedded commands in a special program that could produce quite nicely spaced upper and lower case printing. Access to a data entry service was essential and it was quickly obvious, as the volume of data increased, that adequate data verification techniques had to be built-in. At first, this was by double-entry, which was not satisfactory; followed by data checking on entry, which required some programming to prevent out-of-range data being entered and to ensure the data went into the right columns on the 80-column cards.

From these humble and clumsy beginnings, we move today to a situation where extensive programming is used, optical mark recognition assists some data entry, computer based tests that can record responses directly and be delivered across intranets and internets are becoming essential features of indicator systems. All of this requires that a team of very skilled persons is collected together. We have hired predominantly young scientists and mathematicians, who have, almost without exception, continued to be on a steep learning curve, taking further qualifications, constantly upgrading the work, adjusting the programs, writing software, and making full use of the graphical capabilities now available. Not only is data translated into meaningful sentences contingent upon the data-values, but also into graphs that, for example, change colour when differences are statistically significant.

Without this level of technical expertise, data will not be attractive and easy to access by busy teachers, the turn-around of data will be slow and errors will creep in. Schools want data quickly, within weeks of examination[2] results becoming available. And they want the data 100 per cent accurate. In most research projects, if a small percentage of the data doesn't match, this can simply be reported and ignored, but in indicator systems, schools want every single student accounted for. In consequence, not only must there be teams of expert programmers, but also very high capacity computing equipment. Facilities for printing CDs are helpful, as special, user-friendly software is developed to assist schools in their own explorations of their data. CDs can also be used to deliver computerised tests. The data files returned to schools also need to interface easily with schools' management information software such as timetabling and staff deployment.

If the technical infrastructure is effective, data turn-around quick, data presentation attractive and readily interpreted, then the indicator systems will probably grow and this growth itself demands further technical capabilities, such as running a high-capacity server and creating a central database that can be accessed by researchers and secretaries alike. This central database needs to be relational in order to store efficiently the hundreds of thousands of students with hundreds of variables attached to each student in thousands of schools over many years. It must have an extremely friendly front end, so that secretaries can readily track the mail-out of questionnaires and the return of data, plus a massive invoicing system if individual schools can join the project and pay on their own account. Alternatively, school districts might pay for groups of schools.

Finally the infrastructure needs communication on a regular basis with all schools. Newsletters, a website and conferences are important, particularly as teachers become conference presenters and have a credibility with fellow teachers that researchers lose after some years away from the classroom.

We have been fortunate in working with teachers and headteachers ready to welcome, and make themselves familiar with, streams of data. Some government policies have also helped to make indicator systems important and feasible in the UK: the framework of achievement tests shown in Figure 1, the site-based management legislation requiring school districts to devolve about 80 percent of their budgets to schools, and open enrolment policies allowing parental choice of schools. These were intended both to put schools into competitive situations and also given them some freedom of action derived from having budgetary control.

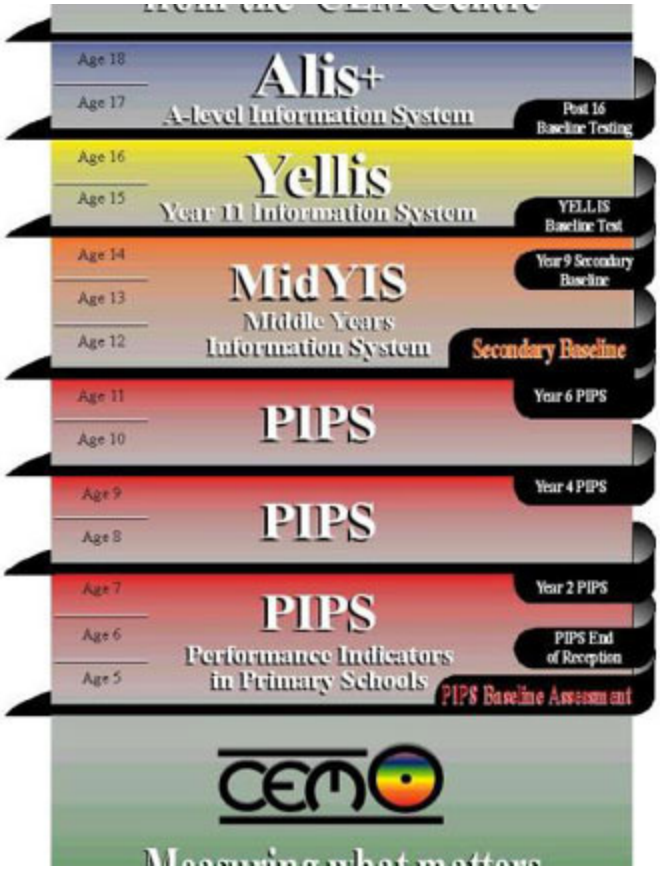

Figure 1. Achievement framework: national tests are provided for students, ages 7, 11, 14, 16, and 18 years.

If the infrastructure for indicator systems can be created, then a cost-effective system is feasible. We now consider the design of such a system, including choosing what to measure, collecting the data, analysing, reporting and interpreting the data.

Choosing indicators

The advice to select a few key indicators is often given (e.g. Lightfoot, 1983 Somekh Convery, Dlaney, Fisher, Gray, Gunn, Henworth and Powell, 1999 p30 and p 34). Whilst this might make life easy, the temptation should be resisted and the advice rejected. A few indicators cannot reflect the complexity of institutions and will undermine the system as gaming takes hold. Given a few indicators, the effort is focused on these concerns alone. Furthermore it is difficult to know which indicators will become important in the future, so that what is now considered to be a key indicator may become of less concern in the future. And who is to decide? Multiple indicators for complex organisations are a fairer representation of the multiple realities within each than is any attempt to assign a single label, whether this label be numerical (e.g. average value added) or verbal (e.g. 'coasting', 'failing').

Our solution is to try to measure what matters as comprehensively as possible. Here the literature in educational research is of value. Bloom's *Taxonomy of Educational Objectives* identified affective, cognitive and psychomotor outcomes that can be taken to include behavioural outcomes. The distinction between aptitudes and achievements (Green, 1974) is an important distinction in the cognitive area. Clearly money matters,

so economic indicators are important. The essence of schooling is who is taught what for how long and by what methods and these concerns can, following the OECD practice (OECD, 1998) be called 'Flow'. Eventually all of these aspects should have indicators. A simple mnemonic makes this list of domains memorable as shown in Figure 2. (See also Fitz-Gibbon and Kochan, 2000.)

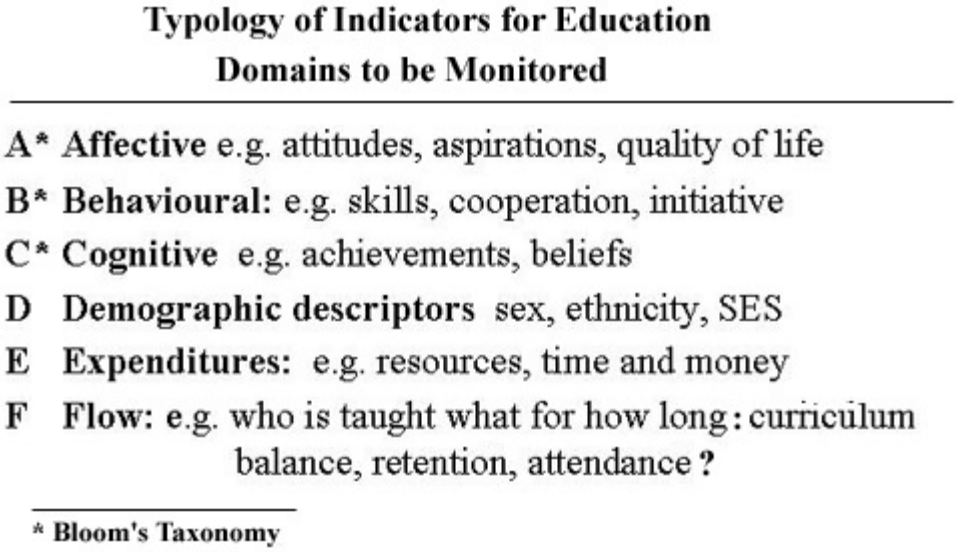

Figure 2. Typology of Education Indicators for Monitoring

The indicators could be collected from various groups such as students, teachers, heads, school districts, states, or parents, the community, the voters. Most indicators could be an input, an output or a long-term outcome and may also be related to a process. Thus a comprehensive classification of indicators can be developed and this also will require a relational database if the measures are to be efficiently stored.

Baseline tests

Prior achievement is an excellent predictor of subsequent achievement, but each student's level of prior achievement will be partly influenced by the effectiveness of the previous stage of schooling. One teacher's output is then another teacher's input. Teachers quite reasonably worry that if they promote high achievement at one age it may be more difficult to show high rates of progress (value added) subsequently. (In Tennessee, they try to control not just for a child's achievement in the previous grade but in the grades two and three years earlier in an attempt to overcome this problem. (Sanders and Horn, 1995).) We have found that the introduction of baseline tests as an alternative to achievement input measures has provided an important alternative approach. The purpose of our baseline tests is not to stamp labels on students, but to predict how easy or difficult it will be to get students through the next set of examinations. What is needed for these tests, then, is *typical* performance, not *maximum* performance. In the secondary school we use tape recordings for test administration so that all schools present students with the same information in the same tone of voice

using exactly the same words and the same timing of each of the subtests. Old IQ tests, in contrast, are often administered in ways that are not properly standardised from school to school.

We have written our own baseline tests to obtain good predictions of subsequent achievement. Our aim has been to obtain quick and efficient measures by using item formats and tasks that require many responses from the pupil that are instantly recorded and therefore add up very quickly to a good predictor of general academic performance. These tests are in some cases remarkably efficient. For example, in 20 minutes the PIPS individually administered adaptive baseline assessment for 4 and 5 year olds obtains measures that predict subsequent progress in mathematics and reading with correlations of about 0.7 (Tymms, 1999). At the secondary school level, our baseline test (the MidYIS test, part of the Middle Years Information System takes 45 minutes of working time and predicts subsequent achievement with correlations of about 0.7, depending upon the subject... such as English or mathematics. The MidYIS test was chosen by the prestigious independent schools for a compulsory baseline.

In addition to prior achievement or baseline measures, are there other important covariates? The best source of information about relevant covariates is not what people write about, but what the data shows. Much is written, for example, of the impact of socio-economic status on achievement, but at the pupil level the correlation is generally about 0.3, thus implying that about 9 per cent of variance in the outcomes will be accounted for by knowing the socio-economic status of the student. In contrast, cognitive measures predict about 50 per cent of subsequent variation. To obtain adequate prediction of subsequent achievement ... and therefore the fairest data for teachers...there is no adequate alternative to a cognitive test.

Affective and social indicators

In addition to the cognitive indicators we need to address the affective and social domains. In Victoria, Australia, there is extensive use of questionnaires to students, to staff in schools and to parents. Currently, in the Curriculum, Evaluation and Management Centre, we concentrate on questionnaires to students, since education is primarily aimed at the students who are in our care for 15,000 hours of compulsory treatment. This concentration on students is also designed to keep the indicator systems lean and efficient and costing as little as possible and obtaining as close as possible to a hundred percent response rates. Students can tell us on questionnaires how much they like school, how much they like an individual subject, whether they feel safe in school, their aspirations for the future, their relationships with teachers, their health, traumas in their lives, how they are taught, and how interesting they find each subject, etc., etc. For children in their first year at school we also ask teachers to rate the children's attention, impulsivity and activity levels.

Does all this amount to too many indicators? Certainly, when schools first join an indicator system, they can feel quite overwhelmed by the amount of data that is returned from a fully developed system. For schools in the first few years of participation 'Keep it simple, stupid' might be a good motto, especially as there is evidence that giving people too much data is de-motivating (Cousins and Leithwood, 1986). However, it would probably be better to give people choice.

We now operate a wide variety of systems of indicators that involve paper- or

computer-based tests, as well as 'basic' or 'extended' versions, the latter including hundreds of variables. We are moving towards systems that will involve on-line administration of data collection and permit matrix sampling and the inclusion of choice, by students or other respondents to questionnaires, of the domains in which they would like to express their views and opinions. This will need close attention to the reliability of the data collected. Thus the matrix sampling will use scales as the unit of sampling rather than items.

Having decided on how to measure the outcomes that matter, one is not finished with the creation of indicators. Just as prior achievement predicts subsequent achievement, so prior attitudes will predict subsequent attitudes, and in order to compare like with like we need to use regression analyses and look at the residuals. The prediction appears not to be so strong as in the cognitive area, perhaps due to less reliable measures, but about 25 percent of the variance of final attitudes in secondary schools is usually predictable from knowing intake attitudes.

Process variables

An indicator system consisting of dependent variables with appropriate covariates is a complete indicator system. However, an indicator system is only a step along the way to trying to understand what works, and how schooling can be improved. Consequently, some of our indicator systems include process variables such as descriptions of methods of teaching and learning for which students in the 16-18 age range report the frequency of use.

Process indicators serve to generate hypotheses and most importantly, they stimulate discussion of teaching methods among staff in schools and as such are valuable. The important problems in trying to attribute cause and effect must, however, be continuously emphasised.

Qualitative data: always valued.

As Berliner (1992) argued, qualitative data are powerful. Early in the ALIS project, one school was constantly at the bottom of the set of participating schools on a scale assessing attitude to school. It paid very little attention to this fact but then open-ended questions were introduced into the data collection and students' comments were typed up and made available to the schools. The typing disguised students' handwriting and kept the feedback anonymous. When the school read statements like 'We are treated like fifth formers without uniform', 'Staff are sarcastic', 'I wish I'd gone to another school' this qualitative data had an impact that was immediate and led to a re-design of the provision for subsequent students. Having had that experience, the school then watched the quantitative attitude indicators with more concern and we continue to provide typed-up responses to open-ended questions.

Credible data collection procedures for attitudinal data

We have already described how the cognitive data collection is standardised so that the same procedures are followed in every school. This standardisation of data collection is important in collecting data that can be validly compared from school to school.

A particular threat to the validity of attitude data could arise from demand characteristics. If students are being asked if they like the school and whether they get on well with teachers yet teachers are looking over students' shoulders, or if the students feel that their questionnaires will be scrutinised by teachers, then the situation becomes subject to possible pressures and influences that could inhibit honest responding. In the secondary school projects, the tape recording that administers the cognitive test introduces the questionnaire part of the data collection by noting that if there is anything they don't understand they should not raise their hand and ask questions because the teacher cannot come to their desk to help them, since the teacher will be staying at the front of the class in order to avoid seeing the responses on any of the questionnaires. Additionally, students are given plastic envelopes in which to seal their questionnaires.

Of course this procedure requires that students can read the questionnaires and this may not always be the case. If there are non-readers, the questionnaire can be tape recorded and students can be given answer sheets with symbols so that they can listen to the questions on the questionnaire and answer on the answer sheet (Fitz-Gibbon, 1985).

Responding to feedback.

The creation of a monitoring system involves a great many decisions and, as a system grows and there is feedback from the users of the system, there is a need to be responsive and flexible whilst holding firm to fundamental principles. In developing an on-entry assessment for 4-5 year olds the intention was that the data would be kept until the children reached the first statutory assessment three years later. But many reception class teachers suggested that we should assess the students again at the end of their first year at school using an extended version of the on-entry assessment. We now do this on a very wide scale and it has proved to be one of the more important innovations with a number of unseen benefits. (For an analysis of the data see Tymms, Merrell and Henderson, 1997).

Matching individual student records from different sources.

The first task in analysing progress data is to match records from baseline tests to outcome measures. The outcome measures should, of course, be curriculum-embedded, high-stakes, authentic tests that reflect work actually taught and worth teaching in the classroom. The use of a standardised multiple choice measure of reading comprehension, for example, is not likely to be fair to schools since teachers may not be able to influence reading comprehension skills once students can read. In other words, there is a problem of lack of sensitivity to instruction. The matching of data from different sources can only be efficiently done by the use of unique identifiers. These preferably should be identifiers containing check digits and the computing facilities to make sure that no identifier is mis-entered.

Transparent analyses vs. sophisticated statistics such as hierarchical linear models.

Einstein said that everything should be as simple as possible, but no simpler. This is a wise, but very challenging, piece of advice. One cannot know how simple a data analysis can be until one has done both simple and complicated analyses and compared the

results with representative sets of real data so that one is looking not only at theoretical models but also at actual magnitudes.

When we won the contract to design a national system of value added indicators, the brief we were given asked for data that was 'statistically valid' and 'readily understood'. These two desiderata could well have been in opposition. We analysed the same data sets using ordinary least squares and multilevel models, and found, as we had found previously, that the average residuals indicating the so-called 'value added' scores for departments or schools, correlated at worst 0.93, and more usually higher, up to 0.99 on the two analyses. Thus it was possible to have the data valid and 'readily understood' by using simple regression. The multi-level analysis, requiring special software and a postgraduate course in statistical analysis, was in contrast to the ordinary least squares analysis that could be taught in primary schools. In our experience in the UK the ordinary least squares analysis can certainly be presented to schools so that most members of staff understand the analysis and can use software to re-analyse data as necessary. This accessibility of the data along with the atmosphere of joint investigation (distributed research) probably helped to encourage acceptance of the indicator systems, unlike the situation that sadly seems to have arisen in Tennessee where a highly ambitious, yearly multi-level analysis was tracking students and teachers (Sanders & Horn, 1995; Baker, Xu & Detch, 1995).

The development of multilevel modelling or hierarchical linear models is admirable, provides efficient calculations and rather different error terms, but to use these procedures in day to day indicator system work is likely to lead to less acceptance of the analysis by teachers. Moreover, it is somewhat akin to applying a correction for relativity when considering the momentum of a moving train: theoretically correct, but in scientific terms, an ill-advised tendency to over-precision.

A recommendation in the Value Added National Project was that prompt initial feedback should be based on very simple value added measures taking account of prior achievement and using ordinary least squares regression methods that any school could adopt and replicate. Then, before any data is made public, statisticians should be given access to the datasets to analyse in numerous sophisticated ways in order to see if any of the analyses makes a difference to particular scores.

Adequate and inadequate statistical modelling.

A method of analysing that *does* make a substantial difference is to consider each subject to have its own regression line, since each subject goes through a particular examining process with a chief examiner and statistical moderation of the marks arrived at by experienced markers working to guidelines. Professor Robin Plackett, winner of two gold medals from the Royal Statistical Society, emphasised in his lectures, usually in his opening sentences that the question to ask, first and foremost, was what processes produced the data. The essence of good statistical modelling is to model the process that produces the data.

From the very start, with the A Level Information System in 1982-83, it was clear that the regression line for mathematics was quite different from the regression line for English and implied that for the same level of prior achievement students would come out with two grades lower taking the Advanced examination in mathematics than they would taking Advanced English (Fitz-Gibbon,1988; Fitz-Gibbon and Vincent, 1994). Regrettably, other researchers (e.g. Donoghue, Thomas, Goldstein, and Knight, 1996) have simply taken the results of all examinations and assumed that the scales could be combined without any

adjustment. Having thus confused the data, sophisticated multilevel models were applied to find that there were differential slopes, i.e. slopes that differed for high and low ability intakes. It was even suggested that teachers may be to blame for concentrating on some groups more than others. This was poor data interpretation since a confound (different subjects with different regression lines) was being attributed to teachers' actions without any corroborating evidence.

In Figure 3, we see some of the different regression segments for different subjects based on intake ranges. These indicate very clearly that the intake differs between subjects, that the difficulty level differs between subjects, and that to simply combine the outcome grades as though each subject were of equivalent difficulty is inconsistent with proper statistical modelling based on the processes that produced the data and that the differences are substantial, unlike the difference made by using or not using hierarchical modelling

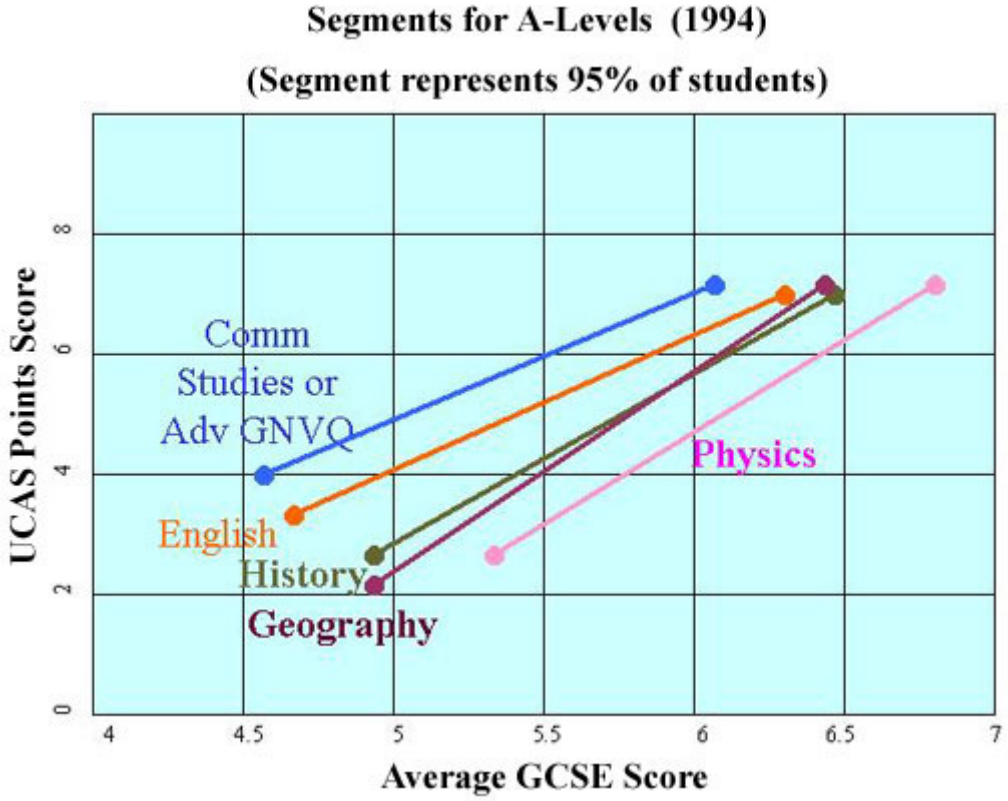

Figure 3. Regression segments showing differences in intake (x-axis) and output (Y-axis) for different subjects

Regression segments, such as were shown in Figure 3 are particularly useful in comparing one subject with another subject, but also in comparing subjects across years. Thus we see in Figure 4 that the average achievement level of the intake is steadily declining (the segment is moving to the left), and the output shows grade inflation (the trend segment is moving up the page). This combination of lower intake range and higher outcome grades has been the pattern with the examinations at age 18 for many years during which time the percentage of students taking these advanced examinations has increased. When these changes are measured against an unchanged baseline, they illustrate the necessary adjustment of 'standards' over time to accommodate expanding range of uptake of advanced courses (Tymms and Fitz-Gibbon, 2001).

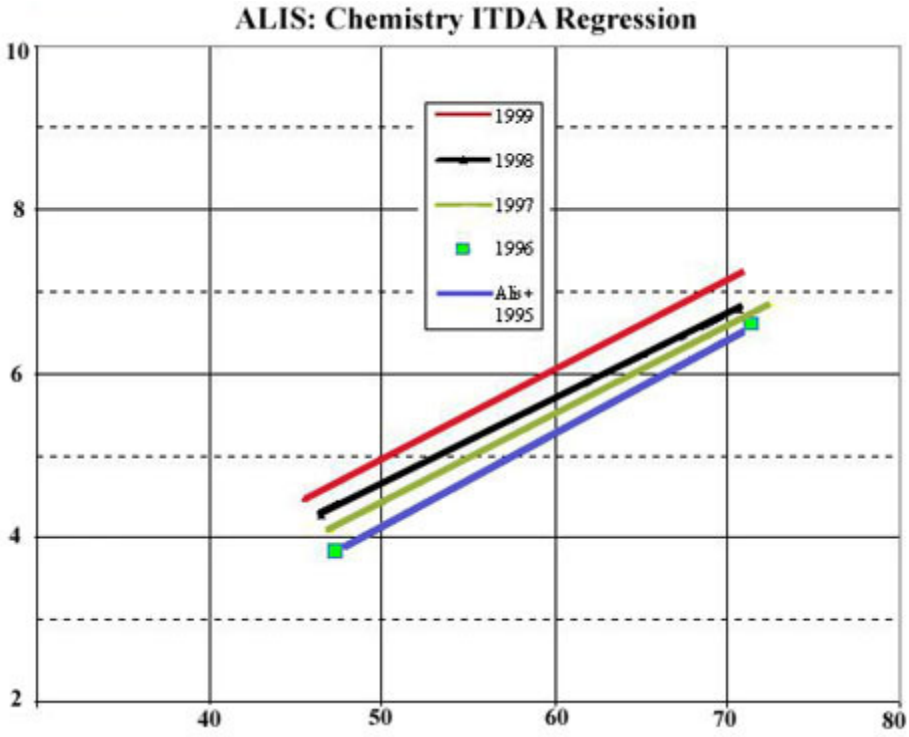

Figure 4. Regression segments for the same subject but different cohorts.

Providing various kinds of feedback, including electronic and web-based feedback.

As with the amount of data, the presentation of data needs to change according to the experience of the school. A school just beginning to get feedback data needs a few clear diagrams and a telephone helpline in case of questions. Schools that have become used to receiving data and have, despite some initial rejection from some departments found it to be useful and credible, start to make more and more use of the data. It therefore becomes valuable to them to have the data provided in Excel spreadsheets, possibly with pre-programmed macros, or in specially prepared software that allows them to undertake procedures such as separating out teaching groups, aggregating by curriculum area, dropping students who have missed substantial amounts of schooling, and adding students for whom data was missing.

Increasingly, as we move from paper-based feedback to sending disks we provide instant feedback. Eventually, with tight encryption techniques this will be directly over the internet.

Chances graphs ... making cognitive tests acceptable.

It has been immensely important in the development of acceptable indicator systems to listen to and to respond to teachers' concerns. It has been important, for example, that baseline tests are not seen as predicting exact outcomes. Fifty per cent of the variation in outcomes is predictable but that means that 50 per cent is not. How can this be represented to teachers who, currently in England, are asked by government agencies to set targets?

This problem was confronted very early in that schools were in some cases preventing students from taking advanced mathematics had they not received a C grade or higher in earlier mathematics courses. When data from a large number of schools was available, in

some of which students had been allowed to take the advanced course even having not done well earlier, it was possible to present what we now call 'Chances graphs' (Fitz-Gibbon, 1992, p.288). These graphs show the chances a student had (in retrospect) of getting each grade subsequently. These 'chances' can be represented with simple bar charts showing the empirical percentages of students who actually achieved each grade the previous year. This empirical distribution has great credibility with teachers and students. It is data that actually happened and if it happened once it can happen again. Thus, the low-achieving student is encouraged to recognise that many low achieving students from the previous year well exceeded the average predicted grade for that starting point. By representing their 'chances', we remove the opposition rightly felt to labelling students with single predicted grades and we provide actual data that is motivating for students.

Statistical Process Control Charts (Shewhart, 1986).

A particularly useful representation of the data is one which answers the question 'How is this department doing from year to year, taking into account the number of students in the group and therefore the expected variation in the average from year to year?' Shewhart's brilliant insight into how to represent confidence intervals has proved most useful. By showing the confidence intervals as guidelines to expected variation, data from year to year are very easily scrutinised. Of course, one expects half the results to be above the line and half below the line in some kind of random order. An example of data from a school that might be concerned about its effectiveness is shown in Figure 5 from the A Level Information system.

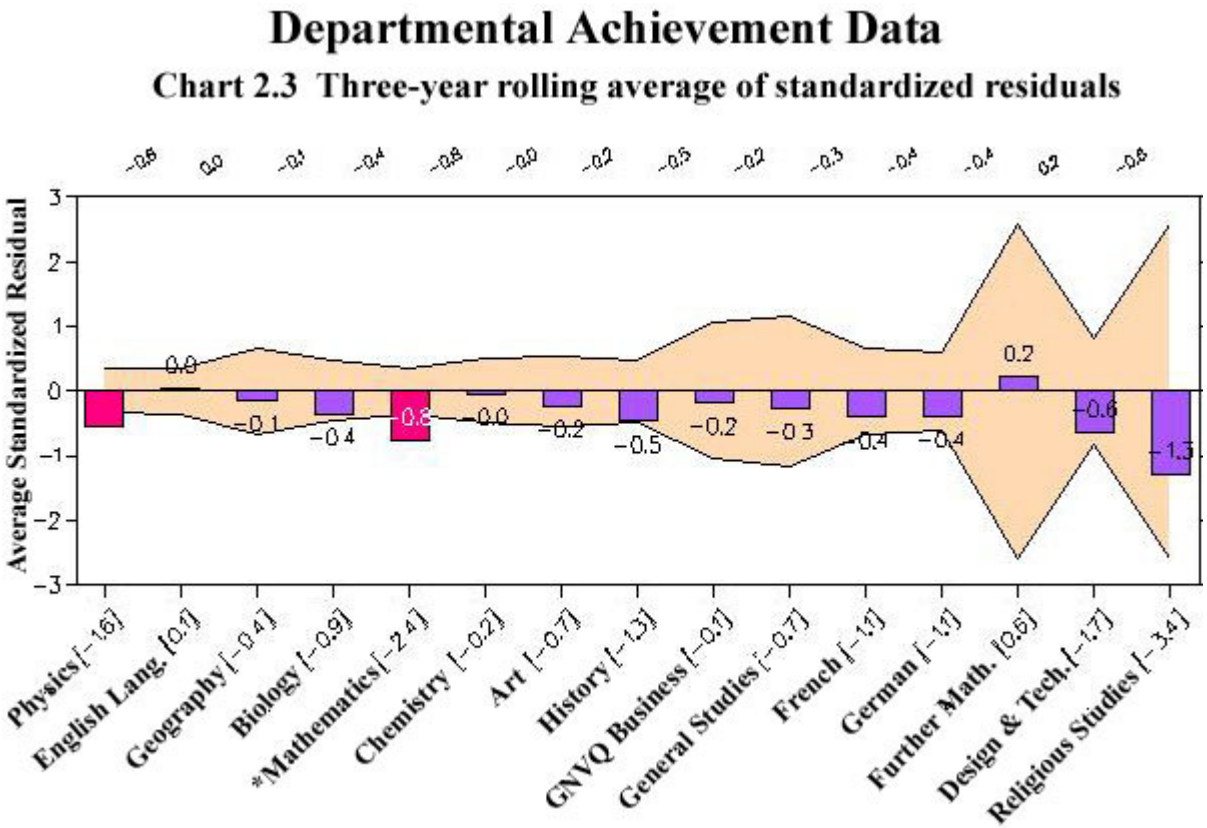

Figure 5. A Statistical Process Control chart for departmental residual gain scores averaged over three years.

The representation involved in statistical process control charts can be applied to presenting

the average residuals from various subjects in the same year or using a three year moving average. Each figure is automatically processed from the data in the relational database and a test for statistical significance made. If the variation from zero is statistically significant, the indicator bar turns red so that schools, at a glance, can see which departments are probably doing better or worse than inherent variation. We also warn, however, that statistical significance at any particular level is not a dichotomy between truth and error but simply an indicator on a continuum. The software we provide enables schools to switch easily between a baseline of prior achievement and a curriculum-free baseline.

For publication: the unit of analysis and the unit of reporting.

Compliance with freedom of information legislation and other relevant laws may require that considerable amounts of data are published. The issue as to what should be published is taken up later since it raises ethical issues.

Let it just be acknowledged here that there are issues regarding the reporting unit (we recommend curriculum area, not whole school nor anything finer-grained) but also the problem arises that the vocabulary of research includes words that raise anxieties such as: 'negative', 'below average' and 'regression'. A solution is to show the data in terms of all-round growth with simply variations in the amount of growth. For lay audiences this representation may be more accessible than regression lines.

Interpreting data: Establishing substantive as opposed to statistical significance.

In the statistical process control charts we saw methods of conveying the inherent variability of data samples. It is highly important that politicians and the public recognise that indicators will fluctuate no matter what teachers do. It was commendable that Scotland waited till it had three years' of data before publishing value added measures.

Although we embed statistical significance tests into the data, we also warn schools against using this as a sole criterion. The problems with routine testing at the 0.05 level have been well rehearsed, (Alkin & Fitz-Gibbon, 1975; Carver, 1975; Glass, McGaw and Smith, 1981; Hedges and Olkin, 1985). To assist schools in interpreting the data, we provide both raw residuals that enable substantive interpretation of differences to be made in the metric in which the examination results are reported[5] and standardised residuals that enable comparisons to be made from year to year. Scales do change; for example in the age 16 examinations, because of grade inflation, an A* was added to the scale as a point above an 'A' grade in the age 16 exams.

Grade inflation due to the standards setting process?

Tymms has suggested that a drift in standards seems to be characteristic in national tests in English primary schools. The reasons for this are connected with the practice of piloting items and setting their difficulty from the results on students who knew they were simply taking an exercise. This 'adrenaline-free' un-prepared testing situation might produce lower performance that would then serve as the benchmark against which the exams were calibrated the following year. Taken under genuine examination conditions, with revision time having been invested and the adrenaline flowing, students might well be producing much better results than those calibrated. Hence the unconscious drift in 'standards'.

Helping users to interpret the data.

It is an unusual teacher-training course that prepares teachers for the kind of information that is now available in England through professional monitoring systems. And yet the information is now seen as vital to many educational professionals. In the course of setting up the CEM Centre projects we have run hundreds of sessions to explain the feedback and to discuss the implications. Further, many courses have been run locally for schools to understand and use the data. Our feeling is that the enormous need for in-service work is an essential part of any monitoring system and that the extent of the need for conferences and workshops often only becomes apparent as the project starts running. It has been standard practice for many years now for our conferences to involve teachers as presenters of the data (e.g. there is a video of a head teacher addressing an early conference and he was a speaker in New Zealand following our involvement there. (Cooper, 1995, video)

Dealing with issues of cause and effect: what works?

This is the most important aspect of data interpretation. It would be wrong to imply to schools that indicator systems are all they need to find out what works. It could take years, even if the search were successful. A school may implement an innovation and the indicator suggests worse results. But perhaps the results would have been even worse without the innovation. Who knows? So the school repeats the innovation and the results stay the same. So another year's data are awaited—and so on.

If instead of this year by year indicator monitoring, if a school joined with 20 other schools and a random 10 implemented one innovation and the other random half implemented a different innovation, all schools would receive 20 years of data in one year. By adopting the methods of science, learning is speeded up and made more reliable.

The fundamental distinction between observation and experimentation must never be blurred.

Epidemiology and clinical trials both have their virtues, but the clinical trials are necessary to establish sound evidence as to what works. That concept applies in education as in medicine and the term 'evidence-based' is now becoming popular. As 'value-added' became the popular word for residuals, evidence-based may become the popular word for experiments. The need not to over-claim for the value of monitoring systems brings U.S. to the next major section of this paper, ethical issues.

Ethical Issues

A major ethical imperative is to do good rather than to do harm. At the very least we might try to observe the Hippocratic oath and 'at least do no harm'. But how do we find out what does harm to students, to society, to academic subjects, to staff?

Evidence of the likely impact of indicator systems on participating schools will be considered including the small number of controlled trials that exist. In addition to this question about the overall impact of indicators there are numerous ethical issues to be addressed that arise in the course of running indicator systems. Each represents a potential source of net harm, a potential negative in a cost benefit analysis.

Some of the questions that arise are:

- Do indicator systems really help schools and affect achievement—or are the admittedly modest funds misspent?
- Should indicator systems lead to a single national, or state, curriculum in order to have a common standard?
- What is the effect of analysing by gender, ethnicity, socio-economic status and religion—does this common activity perpetuate stereotypic thinking?
- What are the effects of poorly chosen indicators, such as those dichotomising continuous data distributions, as in 'percent above x'?
- What are the effects of benchmarking, i.e. comparisons with putatively 'similar' schools?
- Data corruption—does it happen and, if so, who is to blame?
- Is personnel work in public acceptable? (e.g. publishing indicators per teacher)
- Is performance related pay justified?
- Will over-reliance on indicator systems delay the search for better sources of evidence?
- What is the role of the public sector? How can an internal market get the advantages of competition and diversity without the disadvantages of 'the bottom line'? Stakeholders not shareholders?

Do indicator systems really help schools and affect achievement?

It could be argued that because schools freely choose buy into indicator systems this is proof that they find indicator systems useful. However, people buy snake-oil, and the commercial argument is never adequate. People bought treatment with phosphorus that was actually very damaging, and even without a commercial pressure, treatments are provided that do harm simply because adequate evidence has not been collected. What evidence do we have, of a disinterested and objective kind, that indicator systems help schools and, for example, affect achievement?

Cohen (1980) ran a meta-analysis of controlled trials of: no feedback from students to lecturers vs. feedback from students to lecturers vs. feedback from students to lecturers supported by discussions with "an expert." The feedback the same lecturers received in subsequent years improved most in the third condition, and least in the first condition. This result is important. When the ALIS project was about four years old a request was made to a committee at the DfEE (then the Department of Education and Science) inviting them to conduct a randomised controlled trial of the impact of this performance indicator system. The Coopers & Lybrand (1988) report had recommended devolved financing and the use of indictors and the Department was interested. Unfortunately the funds were not found for this potentially important trial. Tymms ran a controlled trial in introducing performance indicators in primary schools into a North Eastern school district in England. A modest effect size (ES_{rct}) of 0.1 was found. This was, however, in 1994 before primary schools were under pressure regarding the publication of examination results and there was no "expert" advice available.

Coe experimented with giving *additional* feedback in the A Level Information System to individual teachers rather than just to school departments. Thus the effect of the randomly assigned feedback was measured not against no feedback but against already substantial feedback, so to expect any further improvement was perhaps optimistic. Nevertheless as a

result of giving classroom by classroom analyses to the teachers concerned, rather than simply departmental data from which this information could be extracted, there was an achievement gain of $ES_{rct} = 0.1$ on the high stakes, externally assessed examinations taken at age 18 years[6]. In the Value Added National Project, Tymms experimented with kinds of feedback and found that for primary teachers, tables appeared to be better understood and also, importantly, appeared to have had more impact than graphical feedback. The average Effect Size across English, mathematics and science was $ES_{rct} = 0.2$ (Tymms, 1997, p12).

In the Years Late Secondary Information System[7], a list of under-aspiring students is produced by combining students' intentions regarding continuing in education with their baseline scores. Many schools given the list of under-aspiring students set up mentoring sessions or special monitoring. Unfortunately, good intentions do not guarantee good outcomes (McCord, 1978; McCord, 1981; Dishion, McCord et al, 1999). Aware of our ethical responsibility not to have teachers wasting their time and in order to avoid harming students, we obtained permission from some schools to only feed back to them a random half of the list of their under-aspiring students. In following up these schools and comparing the outcomes of the named under-aspirers versus the unnamed under-aspirers, we actually have found more differences in favour of the unnamed group than the named group. Indeed, naming students resulted in an overall effect on examination progress, adjusted for prior achievement of value added decrement of $ES_{rct} = -0.38$. Naming seemed to have little effect on whether or not students were counselled at all ($r = 0.01$) but the more counselling sessions that any students, named or not, received the worse were their value added scores ($r = -0.22$). Only 15 schools were involved in this first experiment, but it calls into question many facile beliefs about how achievement can be improved. The findings are challenging and the experiment is being repeated with thirty schools. It illustrates how an indicator system can move the profession forward to proper experimentation.

Should indicator systems lead to a National Curriculum?

The resistance to a National Curriculum in the U.S. has contributed to the slow development of curriculum-embedded, high-stakes, authentic tests. In England, where external curriculum-embedded assessments have been used for decades and school performance tables are published using raw results, moves have been made towards value added systems. These will increase the high stakes nature of the external examinations and, at the same time, government pressure on the Qualifications and Curriculum Authority has led to a reduction from seven independent examination boards to three conglomerates of the former boards. Furthermore there has been a reduction in the number of syllabuses on offer for secondary schools.

Meanwhile in primary schools, a single National Curriculum has been imposed and all primary students sit the same tests designed to the same syllabuses at the ages of 7, 11 and 14 years. The specification of a National Curriculum concentrating on particular subjects and the publication of these data has put schools under pressure to drop attention to such areas as the fine arts, the performing arts, and physical education, and to concentrate on those indicators that are published. All schools are forced to do the same curriculum unless exemptions are granted.

This restriction and concentration certainly represents a downgrading of the professional status of teachers who can now make few important decisions, and it may contribute to

declining levels of satisfaction of teachers. At the very least there should be various curricula available to be chosen, as was the case for decades for teachers of students aged 16 and 18 years. Thus, a teacher who preferred to teach physical geography rather than economic geography could find a syllabus in which the proportion was attractive for that teacher. Another reason for maintaining choice and diversity in syllabuses is that in the entire population a much broader range of skills is thereby likely to be developed. Choice and diversity also keep the examination boards in competition and this ought to lead to an improvement in the quality of the service that they provide. Unfortunately, since they have a virtual monopoly endowed by government approval, it will not be likely that examination boards drop their poor practice unless required to do so. Examples of poor practice from examination boards are leaving the names of students and their schools on the examination paper when it is being assessed. The name of the pupil and the school will often contain clear evidence regarding the pupil's gender, ethnicity, social class and religion. In the face of this information, can essays be read in a totally unbiased way? Further poor practice is the lack of provision of inter-marker reliability data (Fitz-Gibbon, 1996, p.115).

What is the effect of analysing by gender, ethnicity, socio-economic status and religion?

There may be differences between groups, but ethnicity is very poorly defined; socio-economic status is not well-measured; and neither of these variables, is alterable by the school. Alterable variables (Bloom, 1979, 1984) are the key to improvement and accountability. Religion is perhaps an alterable variable, but if we find Catholic schools are doing better than Protestant schools, do we draw the inference that we should make schools turn Catholic? Or vice versa? The habit of analysing by these unalterable variables may simply be a result of the pressure to produce academic papers, whether they contribute to practical or theoretical developments or not. Given a body of data it is easy to break it down by these categories, and report the differences. The fact that it leads nowhere has not been a major consideration in social science research.

The fact that such analyses perpetuate stereotyping should also be a matter of ethical concern. That these correlational analyses do not promote the search for strong evidence as to what works, is certainly a matter for ethical concern. Attention should be directed towards alterable variables rather than unalterable categories into which human beings are grouped, which is the first step to stereotyping. These analyses become particularly a matter of concern when teachers are presumed to be somehow to blame for the 'under-achievement' of boys at the age of 16 as compared with the achievement of girls. Group differences make catchy headlines in the newspapers. While there may sometimes be a need to track group differences, there is a more important need to educate users of data about the size of the effects being studied and what is known about altering the situation. Boys are smaller than girls at age 11. Should they be stretched? Are teachers responsible?

The use of a "percentage greater than" criterion in reporting

The most egregious mistake made in performance data in England has been the DfEE's [8] introduction of arbitrary dichotomies into continuous data. Thus, primary school students' achievements are publicly reported in terms of the percent of students in each school above a certain level, called Level 4. This has the unfortunate implication that students below Level 4 have in some way failed their school or failed in their schooling. This is extremely unethical, since for some students a Level 4 achievement is an excellent achievement, whereas for

others a Level 4 is a failure to reach their potential.

Furthermore, to draw an arbitrary line through a continuous outcome data almost always leads to very negative reactivity. At the secondary level the damaging and unethical impact is a concentration on D students because the reporting line is the percentage of students getting Grade C or above. Time, effort and money have been spent on D students to the neglect of more able and less able students.

If, on the other hand, an average points score is used as the outcome measure, the implication is to work with each pupil to obtain their maximum performance. This is ethical behaviour, it is the kind of behaviour teachers wish to adopt, but it is made impossible by the reporting of indicators based on arbitrary dichotomies in the data.

The effects of arbitrary benchmarks

In England, official bodies such as the Office for Standards in Education, lacking pupil level value added measures, compare schools with 'similar' schools. The classification of 'similar' is usually made on the basis of the percent of students receiving free school meals. However, two schools can both have 20 per cent of students receiving free school meals but otherwise have quite different profiles. For example, one may have a larger proportion of children who also come from schools with very high levels of achievement. Such a school benchmarked against a school with the same percent of free school meals will look very good at the expense of the other school, but the comparison is spurious. Such benchmarking is an inadequate way of making comparisons. The only fair comparisons are with similar students in other schools. There are no similar schools.

It is certainly not ethical to make unfair comparisons which in some cases carry financial consequences for the institution concerned and can lead to job losses and demoralisation. Indeed, to take a most extreme and serious consequence, Ofsted inspectors rely on poor benchmarking data and also sit in classrooms judging teachers. Ofsted inspections have recently been cited in four inquests following suicides by teachers (Times Educational Supplement, April, 2000).

Fair data carefully interpreted is a defence against the inequities of the Ofsted system, problems reported at length to a Select Committee of the House of Commons (website: http://www.cem.dur.ac.uk/) (Kogan, 1999; Fitz-Gibbon, 1998; Fitz-Gibbon and Stephenson, 1999).

Data corruption: when does it happen and who is to blame?

In an article entitled 'On the unintended consequences of publishing performance data in the public sector' Peter Smith, Professor of Economics at the University of York, identified a *'huge number of instances of unintended behavioural consequences of the publication of performance data'* (Smith, 1995). He named eight problems associated with non-effective or counter-productive systems:

- tunnel vision;
- sub-optimisation
- myopia;
- measure fixation;

- gaming;
- ossification;
- misinterpretation;
- misrepresentation.

These can be seen as distortions of behaviour and attention (the first six) and data corruption (the last two). With the sole exception of ossification, every one of these possibilities was raised by headteachers in open-ended items in the questionnaires used in the Value Added National Project. Thus in education these are not theoretical problems but actual, already-perceived problems (Fitz-Gibbon, 1997).

W Edwards Deming (1986) warned that "When there is fear we get the wrong figures." In primary schools in England there have been instances of teachers opening the examination papers the week before assessments and making sure that students were well-prepared. This unfortunately has negative consequences for the school subsequently, since higher than reasonable achievement levels will be expected.

A more subtle form of data corruption is to exclude students who are not going to produce good examination results. In England following the advent of publication of raw achievement levels in the form of 'School Performance Tables',[9] exclusion rates increased 600 per cent. Exclusions from school may be the beginning of an increased risk of delinquency, drug-taking and criminality—is this a price worth paying for the publication of school performance data? It is widely acknowledged that there was a causal link here: schools saw a way to improve their standing in the tables and excluded difficult students. The government some years later responded by publishing exclusion rates and making an issue of 'inclusion'... but the impact had already taken place for many students.

As further pressures arise from 'performance management' (performance related pay systems) it may not be long before we see baseline measures declining so that value added measures look better, particularly when old IQ tests are used for baselines and are not standardised in their administration procedures.

Personnel work in public

Whole school indicators should be avoided because the evidence is that there is more variation within a single school than is generally found between schools. Furthermore, the use of whole school indicators encourages the rank ordering of schools and the public is not prepared to interpret rank orders adequately. Very small differences in the indicator can move a school through many positions in a rank ordering in the middle of a distribution. To avoid simple rank ordering, schools were sometimes put into bands, but this too can be damaging if bands A through E are used. Schools in 'D' and 'E' bands are castigated but in any distribution half have to be below average. This may be politically unpalatable but such is the nature of the average.

If indicators were published for each teacher, this would be tantamount to doing personnel work in public and would be unacceptable. And yet data cannot be withheld from the public unreasonably, so some compromise is needed: not whole school indicators and not individual teacher indicators.

The compromise recommended in the Value Added National Project was to use *curriculum*

area as the unit of reporting. This has the virtue of enabling parents to look for schools that seem to be doing well in the area in which their children are most interested (e.g. performing arts or mathematics and science curriculum areas). Of course, in small schools there may be no distinction between the indicators for a curriculum area and for a teacher. There needs to be some restriction put on the size of sample that can be reported publicly. The CEM Centre is developing these indicators for the provision of data at the LEA[10]/School District level as opposed to individual school level, where the data is presented department by department for affective and cognitive indicators, and student by student in the cognitive area. Within the individual school, further analyses can be undertaken to obtain data teacher by teacher. Such analyses are made easy by our provision of the school's data in software packages called Pupil Assessment and Recording Information System (PARIS).

Performance related pay

George Soros, in his book *The Crisis of Global Capitalism,* elaborates on his concept of reflexivity. His point is that, in the social world, where perceptions can influence behaviour, saying 'it is so' may indeed 'make it so'. Mistaken beliefs about the nature of the physical world have no influence on the physical world, but distortions of beliefs about the social world can have an impact. One of the distortions promulgated by those seeking to implement performance related pay is that pay is the great motivator. This is only a hypothesis, and before huge amounts of money go into implementing performance related pay systems, they should be put to an experimental test in which some schools get performance related pay and other schools get equivalent money to spend as they wish.

The negative influences of performance related pay are potentially the destruction of team work, the demoralisation of those who do not get a performance pay rise, the corruption of data due to the chance to make financial gain from 'good' exam results, and the message sent to students that teachers work for pay: not for their love of the subject, not for their concern for their students, but for pay. According to Soros's concept of reflexivity, this very implication can make itself come true as beliefs can be distorted.

Will over-reliance on indicator systems delay the search for better sources of evidence?

Just as epidemiology is inadequate as a basis for assessing medical treatments, so indicators are inadequate as a means of establishing 'what works' in education. As argued earlier, as schools experience the yearly receipt of indicators of the progress of every student and see the data accumulating in Statistical Process Control charts, they realise that simply watching the indicators, whilst very important, is a slow way to find out 'what works'.

The launch, in Philadelphia in February 2000, of the 'Campbell Collaboration' represents a major effort to create a more just and effective society. It is important that the provision of indicators will support this important step forward and they do, indeed, provide an excellent context in which to conduct experiments: by embedding experiments in institutions with on-going indicator systems, time series data with randomised interventions becomes a very powerful source of high quality evidence.

The role of the public sector

Indicator systems, feasible because of computers, may make the public sector, and in

particular public sector management, a fascinating exercise in applied social science. Finally social scientists may have some responsibility for more than arguments and papers. The actions of managers and administrators should be guided by social science findings. They can study their success in applying the findings by watching the indicators as business managers watch the bottom line or the share price. Perhaps indeed the pensions of Chief Education Officers could be tied to the long-term outcomes of the students who are in their care for about 15,000 hours of compulsory treatment. However, the public sector, including universities, will need to permit innovation, flexibility, and devolved 'site-based management' and public servants will need to reduce drastically time-serving hierarchies and inefficient bureaucracies.

Conclusion

The most important aspect of an indicator system is its reactivity: the impact it has on behaviour in the system being monitored. All the issues raised above need attention to create indicator systems in which the benefits outweigh the costs.

Porter (1988) described the tensions in how indicator systems may be used. When a headteacher[11] said that our indicator systems had 'Introduced a research ethos into the school' we felt this was exactly what was desirable and ethical. But there are pressures to make indicators part of an aggressive management culture, including target setting and performance related pay. Without knowledge of cause, effect and magnitudes of effects this is likely to be unproductive gaming. Good management requires good science, including the recognition of our ignorance concerning many aspects of schooling. An 'Evidence-Based Education Network' is one of the ways in which we wish to promote the research agenda in our 'distributed research' with schools. The questions are not 'Who is to blame and who needs to be rewarded?' but 'What do we know and how do we find out what works?' A research ethos.

Notes

[1]Each summer, with a turn-around time of a few weeks, the CEM Centre processes hundreds of variables and matched pre-post scores on over a million students. Staff look after 12 servers and a relational database management system (RDMS) used by researchers, secretarial and administrative staff.

[2]The examination system in England has long delivered authentic, high stakes, curriculum-embedded tests, called 'examinations'. The complex authentic tests are based on syllabuses to which teachers teach. The examination papers are published each year along with comments from examiners. The systems were set up by universities. Teachers are hired to mark the authentic scripts to clearly designed criteria. The examinations are 'high stakes' but not punitive but aiming to provide certification that assists in gaining university entrance and jobs.

[3]Further discussion of the statistical issues is available in the Vernon Wall lecture on the website www.cem.dur.ac.uk/software/.

[4]Roughly comparable to Advanced Placement in the U.S. Advanced level examinations in England are taken at age 18 and there is, for 2001, a new examination the year before.

[5]E.g. 'Levels' in primary schools and 'grades'... A, B C etc. ... in secondary schools.

[6] (To assist readers in distinguishing correlational from experimental findings the ES is subscripted 'rct' if it arises directly from the manipulation in a randomised controlled trial. This practice, (recommended in Fitz-Gibbon, 1999, p. 37) could make meta analyses considerably easier to conduct, especially for electronically published articles.)

[7]YELSIS also known as YELLIS, Year 11 Information System.

[8]Department for Education and Employment, based in London.

[9]WEBSITE: **Error! Reference source not found.**

[10]Local Education Authority, i.e., School District.

[11]Keith Nancekievil, Gosforth High School, Newcastle upon Tyne, England

References

Alkin, M.C. & Fitz-Gibbon, C. T. (1975) Methods and Theories of Evaluating Programs., *Journal of Research and Development in Education, 8*(3), pp. 2-15.

Baker, A.P., Xu, D. and Detch, E. (1995) The measure of education: a review of the Tennessee value added assessment system. Nashville, TN, Office of Education Accountability, Tennessee Department of Education, Comptroller of the Treasury.

Berliner, D.C. (1992) Telling the stories of educational psychology, *Educational Psychologist, 27*(2), pp. 143-161.

Bloom, B.S. (1979). Alterable variables: the new direction in educational research. Edinburgh: Scottish Council for Research.

Bloom, B.S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as one-to-one Tutoring. *Educational Researcher,* June/July: 4-16.

Carver, R.P. (1975). The Coleman Report: Using inappropriately designed achievement tests. *American Educational Research Journal, 12*(1): 77-86.

Cohen, P.A. (1980). Effectiveness of Student-Rating Feedback for Improving College Instruction: A Meta-analysis of Findings. *Research in Higher Education, 13*(4): 321-341.

Coopers and Lybrand, (1988). Local Management of Schools. London: HMSO.

Cousins, J.B. and Leithwood, K.A. (1986). Current Empirical Research on Evaluation Utilization. *Review of Educational Research, 56*(3): 331-364.

Deming, W.E. (1986). *Out of the Crisis: Quality Productivity and Competitive Position.* Cambridge University Press.

Dishion, T.J., McCord, J. et al. (1999). When Interventions Harm. *American Psychologist,*

54(9): 755-764.

Donoghue, M., Thomas, S., Goldstein, H. and Knight, T. (1996). DfEE Study of Value Added for 16-18 Year Olds in England. London: DfEE.

Fitz-Gibbon, C.T. & Stephenson-Forster, N.J. (1999). Is Ofsted helpful? An evaluation using social science criteria, in C. Cullingford (Ed.), *An Inspector Calls: Ofsted and its effect on school standards* (London, Kogan Page).

Fitz-Gibbon, C.T. (1985). Using audio tapes in questionnaire administration. *Research Intelligence, 19*:8-9.

Fitz-Gibbon, C.T. (1988). Recalculating the standard. 26-8-88, *The Times Educational Supplement*, p 15

Fitz-Gibbon, C.T. (1990). Performance Indicators: a BERA Dialogue. Clevedon, Avon: Multi-lingual Matters.

Fitz-Gibbon, C. T. (1992). The Design of Indicator Systems - The Role of Education in Universities, and the Role of Inspectors/Advisers: A Discussion and a Case Study. *Research Papers in Education—Policy and Practice, 7*(3), 271-300.

Fitz-Gibbon, C.T. (1996). *Monitoring Education: Indicators, Quality and Effectiveness*. London: Cassell/Continuum.

Fitz-Gibbon, C.T. (1997). *The Value Added National Project: Final Report: Feasibility studies for a national system of Value Added indicators.* London: School Curriculum and Assessment Authority.

Fitz-Gibbon, C. T. (1999). Education: High Potential Not Yet Realized*. Public Money & Management: Integrating Theory and Practice in Public Management 19*(1): 33-40.

Fitz-Gibbon, C.T. and Kochan, S. (2000). School Effectiveness and Education Indicators. In C. Teddlie and D. Reynolds (Eds) *The International Handbook of School Effectiveness Research*. London: Falmer Press, pages 257-282.

Fitz-Gibbon, C.T. and Stephenson-Forster, N.J. (1999). Is Ofsted helpful? An evaluation using social science criteria. In C. Cullingford (Ed) *An Inspector Calls: Ofsted and its effect on school standards.* London: Kogan Page, pages 97-118.

Fitz-Gibbon, C.T. and Vincent, L. (1994*) Candidates' Performance in Public Examinations in Mathematics and Science*. London: School Curriculum and Assessment Authority (SCAA).

Glass, G.V., McGaw, B. & Smith, M.L. (1981). *Meta-analysis in social research.* Beverly Hills, CA:, SAGE Publications).

Green, D.R. (1974) *The Aptitude Achievement Distinction*. Monterey, CA: McGraw-Hill.

Hart, P.M., Conn, M. & Carter, N. (1992) *School Organisational Health Questionnaire: Manual.* Melbourne, Victoria, Australia, Department of School Education).

Hedges, L.V. & Olkin, I. (1985) *Statistical methods for meta-analysis.* New York: Academic Press.

Kogan, M. (1999) The Ofsted System of School Inspection: An independent evaluation. A report of a study by The Centre for the Evaluation of Public Policy and Practice and Helix Consulting Group. CEPPP, Brunel University.

Lightfoot, S. L. (1983). *The Good High School: Portraits of Character and Culture*. New York, Basic Books.

McCord, J. (1978). A thirty-year follow-up of treatment effects. *American Psychologist, 33*, 284-289.

McCord, J. (1981). Considerations of Some Effects of a Counseling Program. In S.E. Martin, L.B. Sechrest and R. Redner (Eds.), *New Directions in the Rehabilitation of Criminal Offenders*. Washington, DC: National Academy Press, 393-405.

OECD (1998). *Education at a Glance OECD Indicators*.

Porter, A. (1988). Indicators: Objective Data or Political Tool? *Phi Delta Kappan,* (March 1988): 503-508.

Sanders, W.L. and Horn, S. (1995). An Overview of the Tennessee Value-Added Assessment System (TVAAS)—Answers to Frequently Asked Questions. Knoxville, Tennessee: The University of Tennessee.

Shewhart, W.A. (1986). Statistical Method from the Viewpoint of Quality Control. (Graduate School, Department of Agriculture, Washington, 1939). Dover.

Smith, P. (1995). On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration*. 18(2/3) 277-310

Somekh, B., Convery, A. et al. (1999). Improving College Effectiveness: raising quality and achievement. London: Further Education Development Agency.

Tymms, P.B. (1997). The Value Added National Project – Technical Report: Primary 3, London: SCAA.

Tymms, P.B. (1999). *Baselines Assessment and Monitoring in Primary Schools: Achievements, Attitudes and Value-Added Indicators.* London: David Fulton.

Tymms, P.B. and Fitz-Gibbon, C.T. (2001). Standards, Achievement and Educational Performance, 1976-2001: A Cause for Celebration? In R. Phillips and J. Furlong (Eds.), *Education, Reform and the State: Politics, Policy and Practice, 1976-2001*. London: Routledge.

Tymms, P.B., Merrell, C. and Henderson, B. (1997). The First Year at School: A quantitative investigation of the attainment and progress of pupils. *Educational Evaluation and Research, 3*(2): 101-118.

About the Authors

Professor Carol Taylor Fitz-Gibbon
Director, The Curriculum, Evaluation and Management Centre
University of Durham
England, UK

Email: Carol.Fitz-Gibbon@cem.dur.ac.uk

After 10 years of teaching physics and mathematics in a variety of schools in the U.K. and then the U.S., Carol Fitz-Gibbon conducted a study for the U.S. Office of Education on the identification of mentally gifted, inner-city students and then became a Research Associate for Marvin C. Alkin, at the Center for the Study of Evaluation, UCLA. She completed a Ph.D. in Research Methods and Evaluation, obtained a grant on the design of compensatory education, co-authored a series of textbooks and returned to the UK in 1978 planning to continue work on Cross-age and Peer Tutoring. But the success of an indicator system she developed with 12 schools in 1983 led to other areas. Much of this work is described in the prize-winning book *Monitoring Education: Indicators, Quality and Effectiveness* (1996).

Professor Peter Tymms
Director of the PIPS Project
The Curriculum, Evaluation and Management Centre
University of Durham
England, UK

After taking a degree in natural sciences, Peter Tymms taught in a wide variety of schools from Central Africa to the north-east of England before starting an academic career. He was "Lecturer in Performance Indicators" at Moray House, Edinburgh, before moving to Newcastle University and then to Durham University, where he is presently Professor of Education. His main research interests are in monitoring, assessment, school effectiveness and research methodology. He is Director of the PIPS project within the CEM Centre, which involves monitoring the progress and attitudes of pupils in about 4000 primary schools. He has published many academic articles, and his book *Baseline Assessment and Monitoring in Primary Schools* has recently appeared.

The World Wide Web address for the *Education Policy Analysis Archives* is **epaa.asu.edu**

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

EPAA Editorial Board