DOCUMENT RESUME

ED 379 340 TM 022 725

AUTHOR Fitz-Gibbon, C. T.
TITLE Indicator Systems for School and Teacher Evaluation: Fire-Fighting It Is!
PUB DATE Jul 94
CONTRACT R117Q00047
NOTE 26p.; Paper presented at the Annual National Evaluation Institute of the Center for Research on Educational Accountability and Teacher Evaluation (3rd, Gatlinburg, TN, July 10-15, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Assessment; Educational Improvement; Educational Research; Elementary Secondary Education; *Equal Education; *Evaluation Methods; Foreign Countries; National Competency Tests; *School Effectiveness; *Test Construction
IDENTIFIERS Authentic Assessment; High Stakes Tests; *Indicators; *Monitoring; United Kingdom

ABSTRACT

In 1979, Gene Glass suggested that it might not be possible to evaluate schools nor to create widely applicable research findings, but that the complexity of education was such that merely "fire-fighting," establishing monitoring systems to alert about educational events, was the best approach. In the United Kingdom, monitoring systems are running in one form or another in over 1,000 schools. Two such systems, the A-Level Information System (ALIS) and the Year 11 Information System (YELLIS), are discussed. Both rest on tests and questionnaires administered in schools and the system of curriculum-embedded high stakes authentic testing used in the United Kingdom. Experience with these systems suggests that an external examination system is a fundamental requirements of fair and effective schooling. Without it, neither equity nor effectiveness can be assessed adequately. The external examination system of the United Kingdom is an example of the type of complex system that is likely to lead to the development of self-reliant institutions fed by high quality statistics. Such a system is grounded in a research ethos rather than an evaluative stance. Five figures and two tables illustrate the discussion. (Contains 15 references.) (SLD)

Paper presented at the

CREATE National Evaluation Institute

July 10-15, 1994

Gatlinburg, Tennessee

Center for Research on Educational Accountability
and Teacher Evaluation (CREATE)

The Evaluation Center
Western Michigan University
Kalamazoo, MI 49008

The Institute was supported in part by the Office of Educational Research and Improvement, U.S. Department of Education, (Grant No. R117Q00047). The opinions expressed are those of the authors, and no official support of these positions by the U.S. Department of Education is intended or should be inferred.

Indicator systems for school and teacher evaluation: Fire-fighting it is!

C.T. Fitz-Gibbon
The Curriculum, Evaluation and Management Centre,
University of Newcastle Upon Tyne
England NE1 7RU

Keywords: Monitoring; Indicators; Complexity; Evaluation

ABSTRACT

In 1979, in an article entitled "Policy for the unpredictable", Gene Glass suggested that it might not be possible to evaluate schools nor to create widely applicable research findings. Education was so complex that we might simply have to be content with "fire-fighting": having monitoring systems in place which could alert us when untoward events were happening. This was a prescient article and fits well with modern theories of complexity. The concept of monitoring also fits well with the better manifestations of the methods of Total Quality Management associated with the name of statistician W. Edwards Deming.

In the UK we have created systems which are now running yearly, in one form or another, in over 1,000 schools. There is light but extensive monitoring of the progress made by the some 150,000 students and the systems are growing. One of the several systems that we are developing and running is ALIS - - the A-Level Information System which has been in operation since 1983. Another is YELLIS, the Year 11 Information System which we felt ready to make available nationally in 1994. Both are considered in this paper.

Both ALIS and YELLIS rest on (a) tests and questionnaires administered in schools and (b) the system of *curriculum-embedded, high stakes, authentic testing* which the UK has always known as *"exams"*.

It is suggested that an external examination system is a fundamental requirement for schooling to be fair and effective. It enhances teacher-student relationships, supports delivery standards and underpins quality assurance procedures, in particular by the provision of feedback to the units of respor ibility. Without it neither equity nor effectiveness can be adequately assessed on an on-going basis.

The UK external examination system, a system which has been exported around the world, is commended as meeting many of the concerns currently driving new approaches to assessment. Illustrations are provided of the use of the resulting data in ways which support teacher and school evaluation. The emphasis is on the kind of complex system likely to lead to effectiveness: the development of self-reliant institutions fed by high quality data and developing a "research ethos" rather than an evaluative stance.

Indicator systems for school and teacher evaluation:

Fire-fighting it is!

Unpredictability

In "Policy for the unpredictable" Gene Glass (1979) suggested that it might not be possible to evaluate schools nor to produce widely applicable research findings. Meanwhile, in the last few decades a sea change has swept through science. Physicist P.C.W. Davies has described the change as "nothing less than a brand new start in the description of nature" (Davies, 1987 p 23). From books such as Complexity: the emerging science at the edge of order and chaos (Waldrop, 1992) we gain new insights into how complex systems evolve *in response to feedback from the environment.* Across many disciplines - - -mathematics, physics, evolutionary biology, cell biology, economics, archaeology, computing - - - a coherent set of concepts is emerging. Whether dealing with organisms or organisations, central concepts are those of *feedback* - - - the flow of information and consequences from the environment in which a complex organism is surviving - - - and *local organisation* as opposed to central control. Key factors in evolving systems seem to be the flow and storage of information between its constituent parts, "cells". Complex systems manage to develop effectiveness - - - -as measured by survival and success - - - not by being told what to do but by getting regular feedback and being able to adapt flexibly.

Schools as complex systems

Of course, analogies prove nothing. However, they often underpin scientific intuition and lead to advances. The analogies offered here with regard to complexity theory serve only as a heuristic, a "this might apply" concept. However, since Education is nothing if not a complex system with multiple networks and feedback loops, it would be strange if some of the findings from the new, computer-driven models of complex systems did not apply to Education. (Tymms, 1990; Fitz-Gibbon, 1993; Tymms, 1994)

Whatever the acceptability or otherwise of analogies with developments in science, there is probably widespread agreement on the need for schools and teachers to have good information on their effectiveness (and also efficiency).

We will consider first the cognitive outcomes of education and then consider other important outcomes.

The two systems running most widely at the moment are our ALIS and YELLIS projects, supported by the two UK associations for Head Teachers (National Association of Head Teachers and the Secondary Heads Association.) The major features are summarised in Table 1 and some of these

features are discussed in this paper under the headings "Cognitive Outcomes" and "Affective and Social Outcomes".

COGNITIVE OUTCOMES

The need for feedback

Whereas feedback to teachers on student behaviour in the classroom is immediate and unambiguous - - - with the consequence that most teachers rapidly learn classroom control - - - feedback about their instructional effective is harder to come by. How can teachers learn about the effectiveness of their teaching? How can they know the answer to the crucial question: **are other teachers getting better results in this subject although working with similar students?** The only way to make fair comparisons to provide an answer to this question is on the basis of student-by-student, subject by subject data in the framework of a large scale monitoring system, covering many institutions in which other teachers are teaching the same subjects to similar students. This requires, of course, some agreement on curricula in order to have curriculum-embedded testing. Walker and Schaffarzick (1974) warned adequately against trying to compare outcomes from different curricula.

Providing feedback through a system of external examinations.

In the UK there is agreement on a core curriculum up to the age of 16 years and, crucially, there are externally set and marked examinations at age 16 and at age 18. As a teacher working with students in these year groups you know that you are teaching towards an examination. You have access to examination papers from previous years, you have reports from the examiners and can direct your teaching towards students' performance on these examination.

What are examination boards?

Essentially, Examination Boards are organisations which set and grade examinations. They thus provide independent outcome measures for the cognitive aims of schooling. In the UK they are non-profit organisations. If they make a surplus, it must be ploughed back into the system. They are under the control of the Charity Commission in this regard.

What kind of examinations do they set?

Examinations are set on the basis of published syllabuses which have been developed by working parties consisting of the Board's full time professionals with expertise in the subject area along with representatives of universities, business, industry, teachers and government organisations such as the Department for Education. Questions are written and debated, marking schemes are devised and all this must be accomplished each year and the documents then kept secure until the examination is administered throughout the country.

Although a few questions in some subjects may be multiple choice items the major part of all examinations consists of authentic tasks: writing essays, working out problems, interpreting maps, working with data and so on, as appropriate to the subject. Preparation for such examinations should therefore involve worthwhile activities rather than practice on the tricks of answering multiple choice items at speed.

The likelihood is that these authentic, curriculum-embedded examinations are *sensitive to instructional effects* and can therefore be used to provide feedback to teachers about the effectiveness of the instruction they provide.

How often are these examinations given?

There is a major set of these examinations for 16 year olds (cf. 10 graders in the US). Students may take one or many subjects and the examination is typically 2.5 hours long. Those going on to Higher Education take about 8 subjects at age 16 (These are called GCSE subjects — General Certificate of Secondary Education.) Their average performance on the eight or so subjects provides an excellent index of a student's general academic aptitude. It represents many hours of testing over a three or four week period in May-June, on subjects which have been taught by a variety of teachers.

How are the grades used?

Examination results are sent to schools on a pre-announced day in August and students can collect official certificates. These are, in the UK, fully recognised qualifications widely accepted and credible to: employers, higher education, parents, students, and teachers. Examination certificates are cherished documents and grades provide basic information which are included on CVs.

How are Examination Boards staffed?

By groups of permanently employed educational professionals with a variety of skills: measurement and statistics, subject matter competencies, experience in teaching. These permanent staff are augmented on an as-needed, per diem basis by: practising teachers, representatives of Higher Education, representatives of business and industry and representatives of professional organisations such as teacher unions.

In short, the permanent staff is consistently enhanced by representatives of important constituencies. Democracy, development and responsiveness are built into the procedures — with frequent revisions of syllabuses to reflect the advice of university staff, business, industry, teachers and such others as express their opinions to the Boards.

How do boards function?

The staff, permanent, full-time and part-time, form committees to devise and revise syllabuses, to set and scrutinise examination papers each year, supervise and keep under review the grading standards. An Examination Board undertakes to:

- provide examinations, usually defined by a subject (e.g. Physics) at a given level (e.g. Ordinary level at 16; Advanced Level at 18.)
- publish syllabuses outlining the topics to be examined in each examination arrange for the administration of each examination at sites called Centres
- provide Centres with instructions and materials so that the examinations can be administered under standardised conditions at the same time of the same day throughout the country.
- arrange for the examinations to be assessed by competent professionals
- check the assessments by a variety of statistical measures and sampling e.g. re-marking of borderline scripts; re-marking of scripts awarded a grade more than two grades discrepant from teachers' predictions, re-marking scripts from any marker who is out of line with other markers
- inform students of the results of their examinations on a pre-specified date and provide them with certificates.
- inform institutions of Higher Education of the results when these are needed for admissions purposes
- provide published analyses of the strengths and weaknesses exhibited by candidates in the examinations in each subject each year
- keep in stock previously used examination papers so that these can be purchased and used for practice
- conduct on-going research into the fairness and adequacy of the examinations.

Using examination results in school and teacher evaluation

Given that we have a nation-wide system of external examinations taken at age 16 and age 18 we have there a framework in which the relative progress of students can be measured. The average grade from all the subjects taken at age 16 (the Grade Point Average at GCSE) is the best single predictor for any subject at age 18 (Advanced Level or A-level). By matching the input GPAs with the output A-level grade a scattergram can be plotted and a regression line calculated. Actual data is shown in figures 1 through 5 for Physics, with regressions drawn separately for sex, Examination Board, school Districts (Local Education Authorities) and school departments (i.e. teaching groups for Physics). It can be seen that it is departments which show the greatest range of results and this finding is robust: even with multi-level modelling adjusting for sasmple size. It is departments — a proxy for what goes on in the classroom — which account for the most variation in outcomes. This finding is consistent with a recent attempt to summarise the literature on experimental tests of

"school effects": the meta analysis of Wang, Haertel and Walberg (1993) which supports just this point about proximal classroom vairables as opposed to distal variables:

> "Distal variables, like state, district, and school level policy and demographics, have little influence on school learning. "
>
> Wang, Haertel and Walberg (1994) p. 276

The authors commented that this finding was "inconsistent with current conventional wisdom which argues for policy-driven solutions, like school re-structuring, school -site management, and tougher teacher credential requirements and evaluation." However, we can see here the potential for interactions. School site management without good outcome measures and feedback may be ineffective. Equally, feedback can hardly be effective if given in a situation in which there are few options for action e.g. without the conjunction of site management.

The regression *line* is appropriate in that a line is a good representation. Using a quadratic makes a minimal improvement and makes the method more difficult to describe to audiences for whom any mathematics is a pain. We do not, therefore, use a quadratic except for research work. In fact, for researach work we use multi-level modelling (Rasbash, Prosser & Goldstein 1989; Raudenbush & Bryk, 1986; Aitkin & Longford, 1986). Given the regression line, based on all the data in the monitoring system for the particular subject, a predicted grade can be produced for each student. If many students in a particular class actually achieved higher grades than predicted (positive residuals) this might be due to effective teaching. Thus the average residual for a class is about as fair an indicator as can be produced. It is the best available because the average GCSE score (the GPA at age 16) is reasonably strongly correlated with subsequent achievement at age 18 (correlations from about .5 to .7 are common, across 40 odd subjects).

The average residuals can be expected to vary from year to year and indeed they do. The question then arises as to how large the average residual has to be to merit praise (if the average residual is positive) or concern (if the average residual is negative. The answer to this is that schools will know before the monitoring team will know. We can test for statistical significance and indicate the expected amount of variation from year to year due simply to sampling, but it will be those closest to the data who will arrive at a working knowledge of the *substantive significance* of the indicators.

One way to display the data is by using the format developed by W.A. Shewart and adopted by W. Edwards Deming as a Statistical Process Control graph. Please see (Figure 6: SPC graph) The indicator (in this case the average residual) is plotted from year to year, from left to right There are symmetrical lines above and below the indicator which are called the upper and lower confidence limits. A residual is likely to vary between these limits, from year to year, simply due to the various

samples of students who show up in the class each year. How wide these confidence limits are is determined by three factors:

- how variable the individual results in the class are (the more variation in the class the more the average can be expected to vary and therefore the wider the limits)
- how many students are in the class (the more students the less the averages will vary and therefore the narrower the limits)
- the level of "confidence" chosen, arbitrarily.

This latter may seem unsatisfactory but that is the nature of statistics. (See Carver, 1978)
As already mentioned, practitioners will know before researchers what is substantively important. If, for example, the year 1991 in figure 6, had been a year in which a teacher had been ill, a poorly qualified supply teacher had taken over and students had been very dissatisfied and did worse in Chemistry than in their other subjects, then you may well feel that the limits were well chosen for there had really been a bad set of results that year. If on another occasion you saw such an out-of-limit result, even though you were not aware there had been anything wrong, you might want to investigate. Had the syllabus not been covered? Were the students not taking an important complementary subject? Had the teacher or textbook changed?

This kind of retrospective diagnosis is an important way of learning from experience and developing hypotheses grounded in the data. To assist this process at the level of the individual student we provide, each year, a list of the students showing for each their prior achievement measure, predicted grade, actual grade and residual. Teachers and Heads of Departments study these lists and generate hypotheses as to what might explain them.

Increasing recognition - - but some reservations

There is increasing recognition in the US that Examinations of the UK type are needed. By moving away from secret, multiple choice tests, the scores on which may heavily reflect aptitude rather than school-based achievement, examinations can make a link between effort and achievement clear to students. Below two US educators, Cohen and Resnick, who have a 2.5 million dollar "New Standards Project" which aims to introduce a National Examination Board, put forward something like this view.

> "Today's schooling and testing practices promote the idea that it is native talent or family background that matter, not one's own effort to learn and achieve. Only a few students — those who know early on that they will compete for selective colleges — have any reason to study hard in school. With only few exceptions there is no chance for students to work against a known standard with teachers as their coaches, allies and mentors. We plan to build an examination system in which effort clearly pays off. Students passing a final examination in high school and completing all of their required merit badges will receive a certificate that will signify true accomplishment, not just time in the seat."

> "For an effort-orientated system to work, it is critical that everyone — teachers, principals, parents and students — knows just what is expected of them. So we propose an 'open' examination system, one in which the questions, as well as many responses judged acceptable, are released as soon as the exams are over. The secrecy normally associated with exams would be gone. Students would be working toward a clear objective with clear criteria for success."

However, there are reservations. To save on costs small sample moitoring for the expensive "new" forms of assessment is sometimes proposed. This sounds reasonable since 100 percent samples are not needed to keep a check on outcomes but it misses several points. Small-sample monitoring would not serve the purpose of feedback to teachers. It would be too inaccurate to estimate classroom effects and lacking in interest to teachers who want information on every student they teach, not a small sample. Concern for every student is a vital part of professional motivation and no monitoring system should undermine that. A strategy of going for samples rather than examinations for all, could miss the feedback from authentic tests which is needed first and foremost closest to the chalk face - - - **where change and improvement have to happen.** To get the feedback on effectiveness that will enable each teacher to compare his or her performance with that of other teachers working with similar pupils will require curriculum-embedded externally set and marked examinations for *all* students, and preferably at regular intervals such as ages 7, 11, 14 and 16. At least the costs and benefits should be assessed on a large scale. Such a system will also move towards creating not just "delivery standards" but constantly improving standards.

A further benefit which might accrue could be that such a system will attract and retain the kind of teachers who are motivated by achieving recognisable results and by feedback which reinforces intrinsic motivation. Teachers who attend to measured outcomes may be particularly effective - - - it is at least an hypothesis worth testing.

Furthermore, whilst an examination system may be expensive, pleas of poverty sound odd from one of the richest countries on earth. It depends, surely, on what is at stake. In the absence of controlled experiments designed to create credible cost-benefit analyses we can only guess. Policy experiments should be undertaken by setting up, at multiple sites, the kind of complex examinations which adequately reflect the landscape of learning , examinations which are perceived by students as worth studying for and by teachers as a fair reflection of their efforts.

A final reservation must be that cognitive achievement must not be seen as the single criterion for good schooling. Other outcomes matter. We do not want to set up systems so fraught with anxiety and competition that these destroy the quality of life of teachers and students.

AFFECTIVE AND SOCIAL OUTCOMES OF SCHOOLING

If we care about quality of life, we must measure it. Indicator systems give strong messages about what is considered worth the effort of measuring and this means that we should not restrict any indicator system solely to cognitive outcomes. We must also measure affective and social outcomes. How do we do this?

The simplest way is to use a questionnaire to students. Not, however, a mailed out or handed out questionnaire but one specially administered in school with the aim of getting 100 percent response rate. The conditions of administration must be such that students feel that anonymity is guaranteed and are encouraged to take the questionnaire seriously. We have used various approaches to obtain good quality data. In the early years of the project the questionnaire was administered by the author using a tape recording to ensure the same instructions and explanations were used in each school. Then "data collectors" were trained to take over and a network developed around the country. Examination conditions are set up with desks well separated and staff are requested not to stay in the room or, if they do, not to go near the students. A letter is given to students in advance of the data collection to make it clear that their opinion is being seriously requested.

In recent years some very large schools ("colleges") have chosen to give the questionnaire themselves and in this case other precaution have to be introduced. A data collector from the university is informed of the timing of the administration and may drop in to monitor; students are provided with plastic envelopes which cannot be opened without destroying them and they are asked on the questionnaire, at the end, if the conditions of administration were in accordance with the requirements as listed on the audio-tape, such as having no staff near desks. We also give a phone number for them to contact us should they have any complaints or queries.

Having gone to some lengths to ensure quality in the data, how do we measure attitudes? With a variety of summated items for each of which there is a five point response scale. One set of six items measures response to each subject (e.g. "I look forward to lessons in this subject") Another scale is composed of six items measuring attitude to the school (e.g. I would recommend others to take their A-levels here"). There is also a simple count of a number of extra-mural activities (sports, music, travel) to obtain a measure of participation in variety of experiences, one aspect of "quality of life" perhaps.

An important affective outcome is students' intentions regarding further study, a measure of academic aspirations. The summated scale measuring their "Likelihood of Staying in Education" is regressed against prior achievement so that under-aspiring groups can be identified.

All these measures and more are tabulated or graphed and fed back to schools in individual booklets for each department. This enables the school management to allow the system to be confidential in the early years of participation - - - for indicator systems can appear very threatening until people discover that datasets are not nearly so dramatic nor stark as rumours.

ALTERNATIVES OR SUPPLEMENTS TO INDICATOR SYSTEMS

Quantitative monitoirng is a *sine qua non* but it is not sufficient. Compliance monitoring, unannounced visits, checks on th vliaidy of the data colletion precedures, should also be implemented. In other words a variety of procedures can be adopted to evaluate schools and teachers. Their acceptability will depend upon the use to which the evaluation is put.
There are some methods which are already popular which is not to say that they are effective. Let us consider three such methods: mutual inspections, expert inspections, and market forces.

Mutual inspections

Cannot schools "inspect" or "evaluate" each other? A team of "inspectors" can be created with representatives from several schools for the inspection of a particular school. Then new teams can be formed for the inspection of schools of the erstwhile "inspectors". Private schools seem to favour mutual evaluation. It sounds good - - - evaluation by those actually doing the job, by those who understand. Alas, there are problems with such arrangements. Here is what one participant in a mutual inspection/accreditation visit had to say:

> "Experience over six years of serving on (an accreditation scheme) has demonstrated that the process lacks the rigour necessary to provide schools with constructive criticism. There are no objective criteria in the ... process, and **its value is vitiated by the reluctance of peers to expose themselves to reprisal.** "
>
> Mitchell, R. 1990 Emphasis added.

T e accreditation visitors suffer both from fear of reprisals and from the fact that they are simply pulled away from a demanding job and to fail a school would be to create masses of work for themselves on appeals and recommendations.

> "Perhaps the most damning feature of peer teams is that they visit each others schools. Horse-trading is a constant hidden agenda. You find, after the fact, that punches were pulled at this or that school because the principal is to visit the team chair's school next year. At a particularly poor school, a superintendent refused to go along with my desire to withdraw accreditation because she said: 'I'm not here to deal with the problems of this district'.
>
> Mitchell, 1990, p. 78.

Inspections by experts

In the UK we have a very expensive system of inspections by people who used to be known as HMI "Her Majesty's Inspectors" but who are now working for the Office for Standards in Education (OFSTED). An inspection involves pre-announced visits lasting about a week. Inspectors sit in lessons, interview, and examining documents and then write a report. Some of the reports have designated schools as "Failing" without any numerical data on effectiveness. After more than 100 years of inspection (Matthew Arnold was one of Her Majesty's Inspectors) there have been no studies of reliability, validity or impact, let alone value for money. An inspection is not qualitative research -

--it is too intrusive, disturbing, short term and unsophisticated. Inspection may be simply a remnant of old English power structures, an observation supported by the differences between inspections proposed for a newly independent sector ... the colleges, serving students post 16 years of age and therefore voluntarily in education ... and the school sector. The Further Education Funding Council (for colleges) has published a slim volume about its approaches (FEFC, 1993) with many references to a developing system, a dialogue between assessors and assessed, an encouragement to have quantitative performance indicators and the intention to have a member of staff from the college with the inspectors at all times. Schools on the other hand are told about non-negotiable reports and inspectors who can only be challenged on matters of fact. The legislation for schools specifically refers to "failing schools" and promises to take them over. Further, should schools have data they will be subject to a level 2 fine if they do not show it to the inspectors. (Unused to being treated like criminals, no one in the teaching profession seems to know what a level 2 fine is.)

Market forces

Current government policy is infusedwith a belief in the value of privatisation ...of schools as well as other necessities of life such as water, health care, and electricity. This has led to some actions which have been widely welcomed: the "Local Management of Schools" legislation required school districts to devolve more than 80 percent of their budgets to schools according to a public formula which is driven largely by student numbers. This has introduced considerable competition between schools for students. The intensity of the competition depends upon accidents of bus routes and distances. In some areas it is alleged that the competition lies not in parents choosing schools but in schools choosing parents and their offspring. Studies of the impact of these competitive policies are in progress and the outcomes must be awaited

INDICATOR SYSTEMS AND PUBLIC ACCOUNTABILITY

The viewpoint that indicators are to be used for feedback which will motivate and assist in improvements can only be part of the scenario. The issue must be joined as to other uses to which indicators can be put. In the UK for example it is only a matter of a short time before there are national regression lines available and then inspectors will be able to compute residuals within schools. This is probably helpful in that data has a moderating influence. Should the school's residuals be publicly available? Should they be published in newspapers? The position that we have adopted in our systems is that since the data show most of the variation to be at the classroom level whole school indicators are of limited usefulness. They might be published as a general re-assurance and check for outliers but they are unlikely to be helpful in, for example, parents' selection of schools. To use whole school indicators is to imply that schools vary as *a whole.* Yet given the variation within schools it is within schools that quality must be assured and improvements made. Should indicators, then, be published department by department? Since departments may sometimes

represent one or two staff, such publication would be tantamount to publishing personnel work. No profession actually undertakes personnel work in public.

As in all complex systems, we have to ask what system will produce the best outcomes, not what conforms to a feel-good agenda.

I am not sure that the answers to the issues raised here can be worked out. What is clearly needed is some long-haul (5 to 10 year) multi -site field trials. In the US there is a chance to create such experiments. In the UK the time has probably passed for 16 to 18 year olds since there will shortly be nationally available data and how it is used will be decided more by political processes than by professional or scientific concerns.

FUTURE DEVELOPMENTS

Going beyond Value Added

There is currently, around the world, concern with the "quality" of schooling and this is translated into a concern about student achievement - - -how good are the grades? But is it not equally important to ask "*Grades in what* ? Does not *what is studied* have at least as much impact as *how effectively it is studied* ? Is most of the research on school effectiveness focusing on short-term, immediate impacts which subsequently have little effect? A grade "A" or a grade "C" might not be as important for subsequent usefulness as the nature of the subject which has been studied. School effectiveness studies should give greater consideration to measures of 'yield' in various areas of the curriculum rather than the current emphasis on relative performance or "value added" . (Walker & Schaffarzick, 1974; Preece, 1983)

CONCLUSIONS

Is it possible to evaluate schools? Yes, to a degree. If you are willing to attribute to schools the responsibility for their residuals (measures of relative pupil progress) then numerical measures can be developed within a framework of fair, curriculum -embedded examinations. Whole-school indicators are then available. However, it is likely to be more profitable to give detailed data to schools, regularly. The evidence is that the major source of variation is what happens in classrooms, not the organisational or structural features of schools and, given this situation, whole school indicators are not useful. To use some of the language of complexity theories, self-organising classrooms need a constant flow of valid information and energy (resources) so that they can find their own ways to thrive on the edge of chaos.

Whilst independent, external, professionally-run examination boards need to be developed at State level, such systems could start on a small scale, even among teachers in just one subject area in one School District.

The necessary framework of fairness based on external marking of students' achievements can be expected to provide motivation for students and to enhance teacher-student relationships. Attention to affective and behavioural outcomes no less than cognitive outcomes will give the right messages

and may help to promote action-research based on data as an on-going way to manage schools for effectiveness in diverse, evolving ways.

Alternative systems designed to insure quality and create improvements are in operation. These include certification programs often favoured by private schools and, for state schools, "inspections" i.e. panels of visitors funded by government. The deficiencies of these systems were discussed.

Notions such as those of unpredictable systems, non-linear dynamics, cellular automata, emergence and efficiency on the edge of chaos now give us a new set of mental models which will without doubt impact on social science.

> "... in the poorly defined, constantly changing environments faced by living systems.....**there seems to be only one way to proceed: trial and error**, also known as Darwinian natural selection."

Thus, Waldrop quotes Langton (p 282) and Langton echoes a sentiment that Karl Popper would recognise. Popper, the scientists' philosopher, wrote of the unpredictability of the future and the need to strengthen our information systems and move forward slowly with evaluation, to undertake 'piecemeal social engineering'. H. A. Simon wrote of the need to try out designs, of the impossibility of predicting which designs would work and he emphasised the difference between this gradual pushing of the boundaries forward to attain certain goals and the investigation of lawful behaviour in nature: The difference between **design and science**. This all fits in with Gene Glass' s conclusion, based on years of looking at the results of educational interventions: that there are unlikely to be general rules.

Nevertheless, whilst the weather is unpredictable the climate is not. There is both complexity and simplicity and it seems that what is needed in the next decade is the development of a variety of indicator systems run in a variety of ways. Feedback of valid analyses of results, to the units of action, on all those outcomes that we care about enough to measures.

REFERENCES

Aitkin, M and Longford, N. (1986) 'Statistical modelling issues in school effectiveness studies', Journal of the Royal Statistical Society, Seriees A, 149 (1), pp 1–43.

Carver, R.P. (1978) 'The case against statistical significance testing' Harvard Educational Review, 48 (3), pp 378–399.

Davies, P.C.W., (1987) The Cosmic Blueprint, London: Heinemann.

Fitz-Gibbon, C.T. (1993) Monitoring School Effectiveness: Simplicity and Complexity. Paper presented at ESRC Seminar Series in School Effectiveness and School Improvement, Sheffield.

Further Education Funding Council (1993) Assessing Achievement. Circular 93/28: Coventry.

Glass, G.V. (1979) Policy for the unpredictable (Uncertainty Research and Policy). Educational Researcher 8 (9) pp 12-14.

Mitchell, R. (1990) 'Site Visits in the Accreditation Process of the Western Association of Schools and Colleges (WASC)' Evaluation and Research in Education 4 (2), pp 75–80.

Preece, P.F.W. (1983) 'The qualitative principle of teaching' Science Education 67, pp 69–73.

Rasbash, J., Prosser, R. and Goldstein, H. (1989) ML3 Software for three level analysis, London: Institute of Education.

Raudenbush, S. and Bryk, A.S. (1986) A hierarchical model for studying school effects, pp 1–17.

Tymms, P.B. (1994) Theories, Models and Simulation: School effectiveness at an impasse, Paper presented to the ESRC School Effectiveness and School Improvement seminar: London.

Tymms, P.B., (1990) 'Can Indicator Systems Improve the Effectiveness of Science and Mathematics Education? The Case of the UK' Evaluation and Research in Education 4 (2), pp 61–70.

Waldrop, M.M. (1992) Complexity: the emerging science at the edge of order and chaos. London: Viking.

Walker, D.F. and Schaffarzick, J. (1974) Comparing Curricula — Review of Educational Research, pp 83–112.

Wang, M.C., Haertel, G.D. and Walberg, H.J. (1993) 'Toward a knowledge base for school learning', Review of Educational Research, 63 (3), pp 249–294.

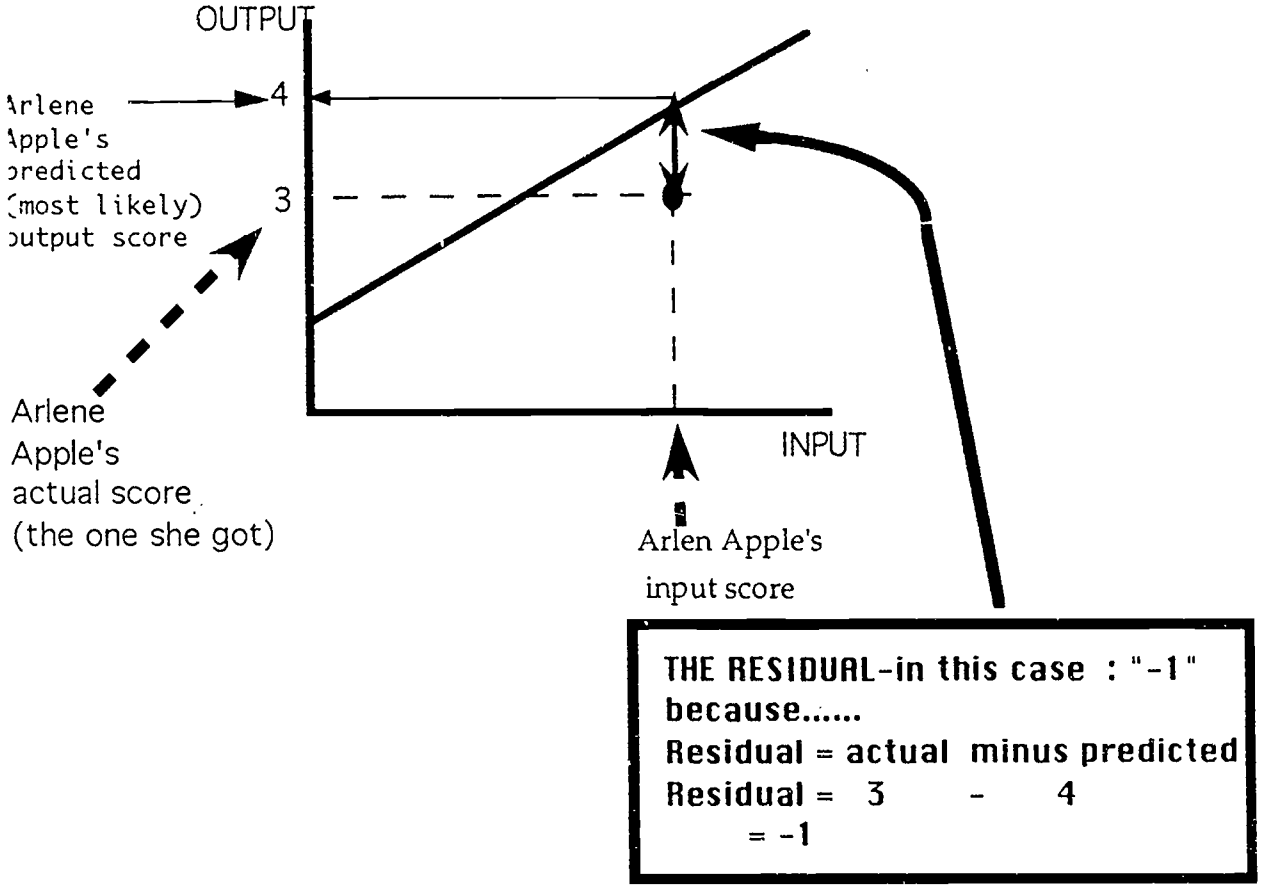


FIGURE 0. A residual: the difference between the statistically predicted grade and the actual grade achieved.

FIGURE 6. Shewart-style graph for monitoring school effectiveness.

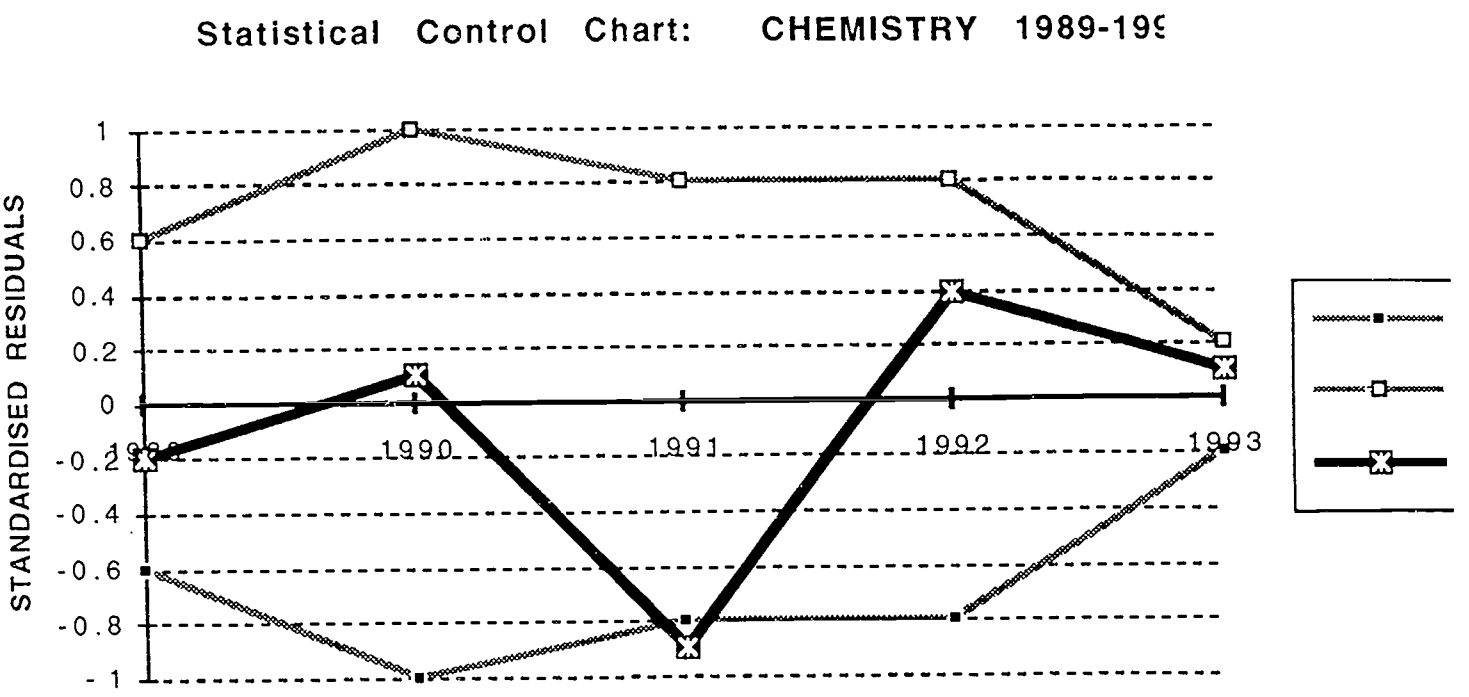
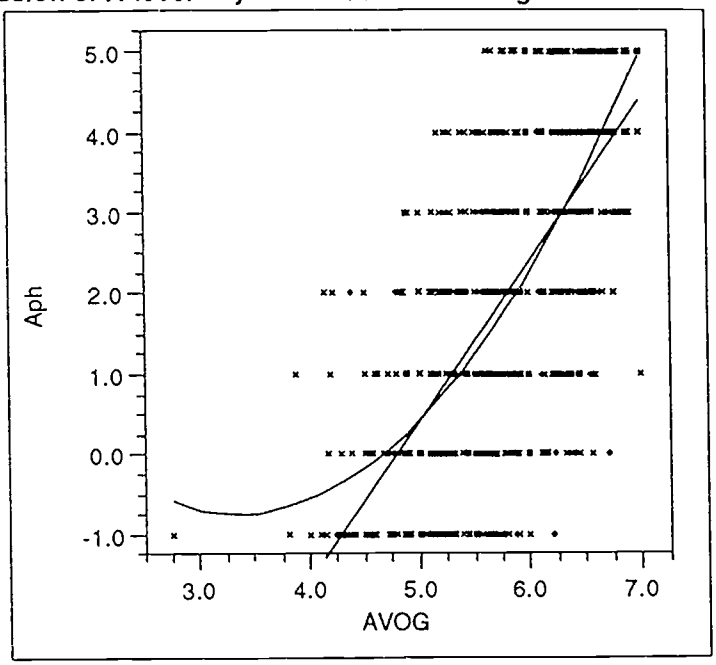


FIGURE 1 Regression of A-level Physics Grades on average examination grades at age 16



Notes:

1. The vertical axis represents grades on a single subject taken at age 18, in this case Physics. The horizontal axis represents the average grade obtained on all examinations taken at age 16 (usually about 8 subjects).
2. For "Aph" i.e. the Advanced-level grade in Physics,

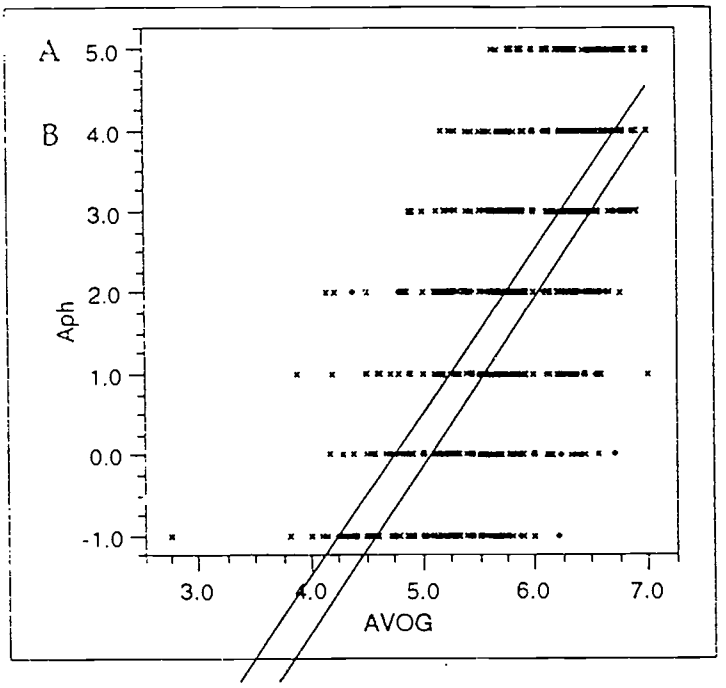
   5=A, 4=B etc.

   . For "AVOG" i.e. Average grade at age 16,

   7=A, 6=B etc.

   These particular scales are used for historical reasons
3. A linear and a quadratic equation have been fitted, giving the line and the curve.

FIGURE 2 Regression of A-level Physics Grades on average examination grades at age 16, *with separate regression lines by sex*



SEX=1 (male)

Summary of Fit

Rsquare	0.47
Root Mean Square Error	1.43
Mean of Response	2.06
Observations (or Sum Wgts)	693

SEX=2 (female)

Summary of Fit

Rsquare	0.49
Root Mean Square Error	1.32
Mean of Response	1.85
Observations (or Sum Wgts)	189

FIGURE 3 Regression of A-level Physics Grades on average examination grades at age 16:

by Examination Board

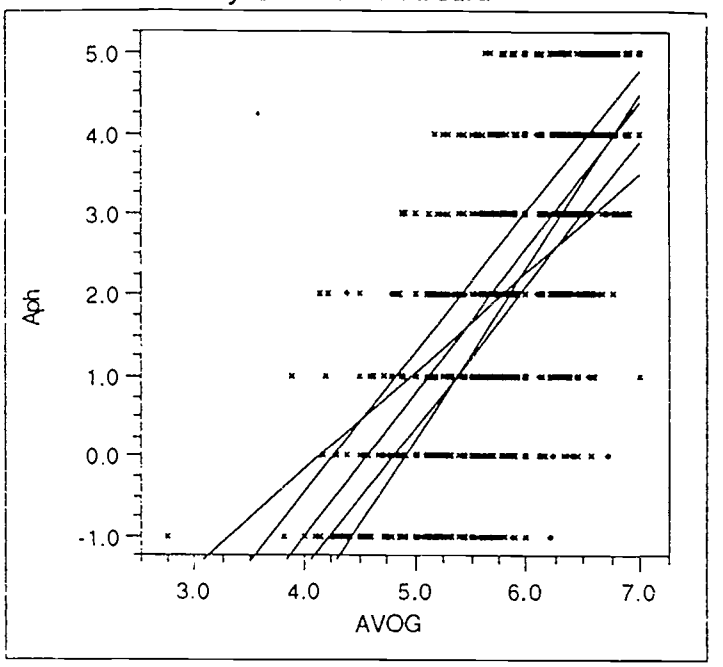


FIGURE 4 Regression of A-level Physics Grades on average examination grades at age 16:

by school district (LEA)

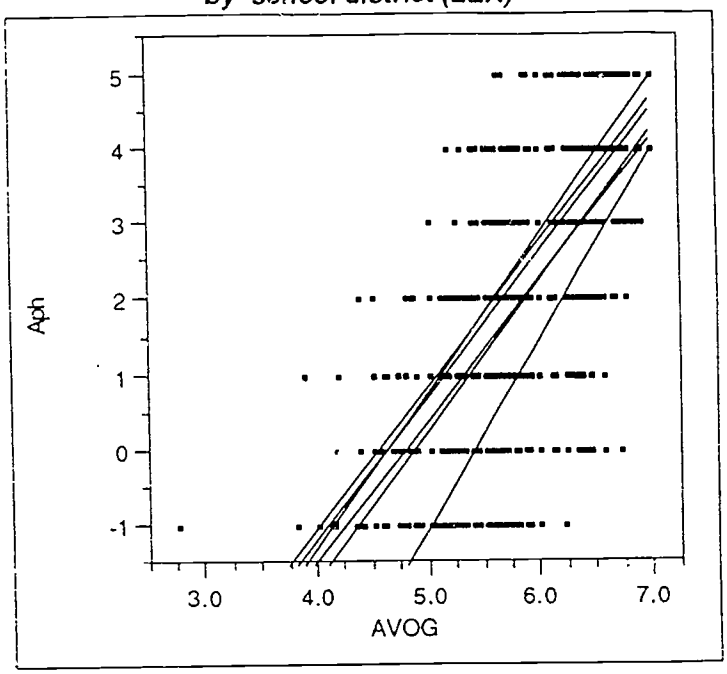
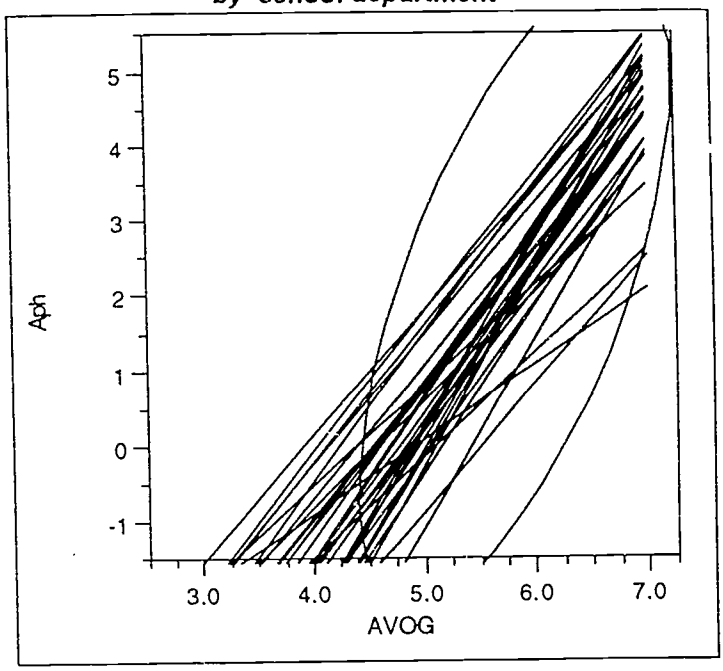


FIGURE 5 Regression of A-level Physics Grades on average examination grades at age 16:

by school department



Variable	Mean	Std Dev	Correlation	Signif. Prob	Number
AVOG	5.826617	0.661413	0.67	0.0000	665
Aph	2.090226	1.950667			

Table 2 Correlations

	A-level Physics grade	Average Ach. .@16	SEX	Attitude to school	Attitude to Physics	Residual Physics
A-level Physics grade	1.00	0.67	-0.05	0.15	0.33	0.74
Average Ach..@16	0.67	1.00	0.12	0.11	0.18	0.02
SEX	-0.05	0.12	1.00	0.06	-0.06	-0.17
Attitude toSchool	0.15	0.11	0.06	1.00	0.32	0.09
Attitude to Physics	0.33	0.18	-0.06	0.32	1.00	0.28
Residual Physics	0.74	0.02	-0.17	0.09	0.28	1.00

Table 1. Indicator Systems for schooling effects on 16 and 18 year olds.

OUTCOMES	Indicator System For Students' Progress Between The Ages Of:	
.	14-18 years(the YELLIS project)	16-18 years (The ALIS project)
Cognitive outcomes	Students' achievement in external examinations at age 16 (General Certificate of Secondary Education (GCSE)	• students' achievement in external examinations age 18 (A levels)*
Affective outcomes	• students' attitudes to main subjects: English, Math. & science • students' attitudes to their school or college • students' response to school processes • students' feelings of being over-or under-challenged in the major subjects	• students' attitudes to each subject studied for the examinations at age 18 • students' attitudes to their school or college ("customer satisfaction")
social outcomes	• students' aspirations • students' freedom from fear (bullying, insults, safety)	• students' aspirations • students' participation in extra mural activities (as a 'quality of life' indicator)
Behavioural /life chance outcomes	• continuing in education after the period of compulsory schooling • jobs achieved	• earnings • perceived quality of life • retrospective satisfaction • residual effects effective schooling on achievement in higher education

*The externally set and marked examinations would be described in the US as "authentic" (essays are written, problems are worked out, maps are read, data is analysed), high stakes (certificates are issued which influence admissions to further educational opportunities and job prospects) and curriculum embedded (the examinations test what has been taught, which has followed published syllabus).

List of figures and tables.