Curriculum, Evaluation and Management Centre

PUBLICATION NO.04

THE DESIGN OF INDICATOR SYSTEMS, THE ROLE OF EDUCATION IN UNIVERSITIES, AND THE ROLE OF INSPECTORS/ADVISERS: A DISCUSSION AND A CASE STUDY

C.T. FITZ-GIBBON

Oxford Review of Education

The design of indicator systems, the role of education in universities, and the role of inspectors/advisers: a discussion and a case study

Carol Taylor Fitz-Gibbon

Abstract

Herbert A. Simon's concepts of the design of complex systems are used as a framework in which to describe the development of an indicator system. The implications for Education as a discipline and for the structure of the education service are considered. Propositions include the need to consider *design* in contradistinction to *science* and the advisability of reconceptualizing the roles of educational inspectors. As educational decisions are devolved to school sites, inspectors have one of the few educational roles left in local education authorities: monitoring schools, assessing effectiveness. It is a role which can develop fruitfully if regarded as encompassing mainly the provision of feedback on outcomes and the collaborative design of systems that work.

Reactivity (behavioural impact) should be the prime criterion when indicator systems are evaluated in terms of the standard canons of research practice.

Introduction

- What does a successful performance monitoring system look like?
- Should education exist as a discipline in universities? Is it a social *science*?
- Is there a role for local education authorities (LEAs) or should they, as has been suggested, 'wither on the vine'?

These questions are not unrelated. They are considered in this article by using, as an example, a performance indicator system designed to monitor A-level provision.[1] ALIS, the A-Level Information System, started in the academic year 1982–83 with a dozen schools. It has recently grown rapidly, with more LEAs joining and some individual schools and colleges choosing to participate at their own expense (Fig. 1). The genesis and features

Carol Taylor Fitz-Gibbon is Professor in the School of Education, University of Newcastle-upon-Tyne.

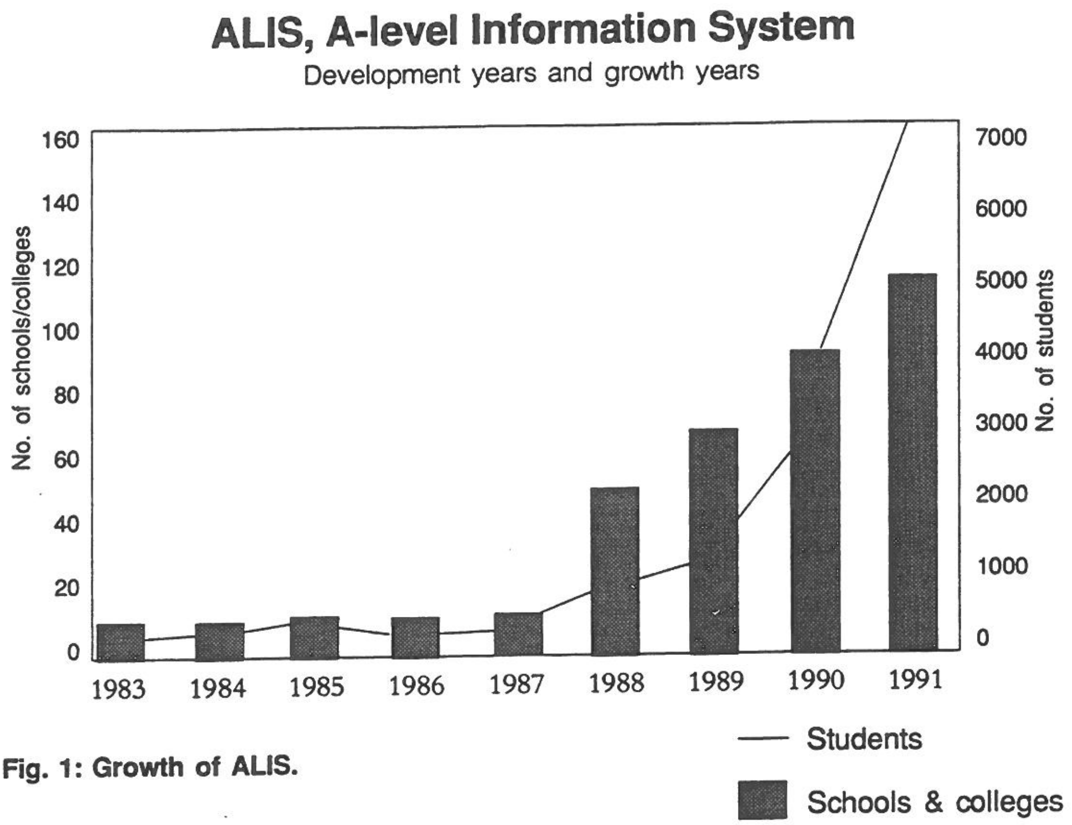


Fig. 1: Growth of ALIS.

of this system will be described in the framework of a concept which has implications for answers to all three questions posed above: the concept of *design* as put forward by H. A. Simon.

The concept of design

Herbert A. Simon is Professor of Computing and Psychology at Carnegie-Mellon University in the United States. He was awarded a Nobel prize for his work in economics and has had substantial impact on organization theory and concepts in artificial intelligence. Much of his thinking about complex systems — and the state education system is one such complex system — was presented in a book which has become something of a classic: *Sciences of the Artificial* (Simon, 1988; first edition, 1969.)

The artificial environment: the man-made world of artefacts

We live in an environment of artefacts: of letters, of words, of television screens, of cars, of computers, of examination results, of certificates. We have created an artificial environment of objects and of symbols. These objects and symbols are there to serve purposes. They have been designed to achieve goals. The creation of many of these artefacts rests on scientific knowledge but the process of creation is not a pre-specified nomothetic science but, rather, a problem of design.

> Certain phenomena are 'artificial' in a very specific sense: they are as they are only because of a system's being moulded, by goals or purposes, to the environment in which it lives. If natural phenomena have an air of 'necessity' about them in their subservience to natural law, artificial phenomena have an air of 'contingency' in their malleability by environment.
>
> The contingency of artificial phenomena has always created doubts as to whether they fall properly within the compass of science. (Simon, 1988; p. x)

Simon argues that the professional schools, such as schools of education, law and medicine, '... are concerned not with the necessary, but with the contingent — not with how things are but with how they might be — in short, with design'. (*ibid.* p. xi)

> Everyone designs who devises courses of action aimed at changing existing situations into preferred ones. The intellectual activity that produces material artefacts is no different fundamentally from the one that prescribes remedies for a sick patient or the one that designs a new sales plan for a company or a social welfare policy for a State. Design, so construed, is the core of all professional training; it is the principal mark that distinguishes the professions from the sciences. Schools of Engineering, as well as Schools of Architecture, Business, Education, Law and Medicine are all centrally concerned with the process of design. (*ibid.* p. 129)

Despite the fact that *Sciences of the Artificial* was first published in 1969, this concept of design, as distinct from science, has not yet become widely embraced and there are still signs of identity problems in professional 'schools', including university departments of education (UDEs) (Clifford and Guthrie 1988; Thomas 1990). There is also what might be termed an imposed identity crisis for inspectors and advisers.[2] Their independent funding, allowing them to speak 'without fear or favour', is being largely withdrawn.

Having mapped out three of the major concerns of this document, educational activity as design, the role of UDEs and the role of LEA inspector/advisers, let us look more closely at design concepts and follow them in the case of an indicator system.

Design: what is known about the process?

Simon describes the design process and its relationship to science. Design is the utilization of scientific knowledge to create an artefact *for a purpose*. It is neither pure research nor applied research but a creative process directed by purposes. To achieve the purposes, two major aspects must be researched and taken into account: the outer environment in which the artefact must exist (the context), and the inner mechanisms or processes involved in the artefact itself.

Most artefacts created in the social disciplines involve complex systems and Simon discusses at length the features of such systems. Complex systems, he argues, are almost inevitably *hierarchical*, working at various levels. This is certainly true in education and this multilevel structure is an important reason why multilevel modelling (Aitkin and Longford, 1986; Raudenbush and Bryk, 1986; Goldstein, 1987) has become so important in the analysis of school effectiveness. The education system consists of nested hierarchies: the pupil, the teacher, the classroom, the department, the school, the LEA, the nation. The constituent units of hierarchies need to be semi-autonomous modules in the complex structure.

Simon notes that in analysing an organization the decomposition into units will not be a unique one:

> To design ... a complex structure, one powerful technique is to discover viable ways of decomposing it into semi-independent components corresponding to its many functional parts ... there is no reason to expect that the decomposition of the complete design into functional components will be unique. (*ibid.*, pp. 148–9)

The education system itself is a complex system but this paper is concerned not with the design of the education system but with the design of a monitoring or feedback system for education, that is, with the design of a component of the larger system.

In designing indicator systems the approach of 'modularizing' will undoubtedly be needed. For example, the ALIS indicator system represents just one module of what must eventually become many interlocking modules providing indicator systems. The ALIS project deals with the module of the education system consisting of students opting to spend 1 or¿2 years studying for the A-level examinations, examinations which form a major hurdle to be overcome in reaching higher education and the professions. Subsequently, monitoring systems will be needed for other modules: for the two years leading up to GCSE,[3] for example. And when each of the modules works then other modules must be designed to look at the flow between modules.

Designs, however, do not emerge finished, born effortlessly from first thoughts. Computer programs, simpler than social systems, need to be debugged and so too do monitoring modules. Designs require what Simon called the generator–test cycle:

> ... think of the design process as involving, first, the generation of alternatives and, then, the testing of these alternatives against a whole array of requirements and constraints. (*ibid.*, p. 149)

The requirements and constraints arise from the nature of the goals and the environments in which the design must work. Simon conceptualizes design as the creation of an artefact which has a certain 'inner environment'. The inner environment is the way the artefact functions to achieve goals. The artefact must exist, however, in an 'outer environment' and the goals will only be achieved as the inner and the outer environment interact. Design, then, may be conceived as an interface problem between an inner and an outer environment. Putting these concepts into educational terms, the outer environment is represented by the constraints of resources and the impact of the policies and *zeitgeist*, along with the existing systems such as examination boards and university entrance policies. The inner environment will be the techniques available to create indicators. The design problem is to create a monitoring system which works.

In short, the design process will be driven by *purposes or goals*, it will draw on a *knowledge base*; it will often require *modular approaches* and will involve the methods of science in testing the design. Finally it must be recognized that design is an *on-going process*. Simon speaks of the ongoing nature of design, of design as a creative activity in which many persons participate. There is no end to the design process: 'The idea of final goals is inconsistent with our limited ability to foretell or determine the future' (*ibid.*, p. 187).

The application of these ideas to the design of the monitoring system is discussed in the following section.

The design of the monitoring system

The *purpose* of monitoring is to improve education. Whether in fact a monitoring system *will* improve education cannot be known *a priori*. The system will have to be assessed as part of the generator–test cycle referred to above. We need also to consider the feasibility of alternatives. The very attempt to create such a project as ALIS implies either the lack of existing alternatives or some dissatisfaction with existing alternatives.

What are the alternatives to the approach of formal, quantitative monitoring? If one wants to know whether or not schools or teachers or departments are effective, can we not just ask those who have close, first-hand knowledge of these teachers, departments or schools? Can we not ask inspectors, who visit many of these teachers, departments and schools? Can we not just leave it to headteachers? In short, can we not just ask 'experts'?

Despite the willingness with which some will lay claim to expertise, many of Her Majesty's Inspectors (HMI), local inspectors, advisors and headteachers will agree that it is not easy to know with any confidence how effective individual teachers or school departments are. Certain styles may be *preferred* — by some — certain individuals may be *liked*, but are they *effective* in, say, preparing pupils for examinations or creating a love-of-the-subject in pupils?

An apparently good set of examination results often leaves everyone wondering whether the results should actually have been even better or could reasonably be considered to be exceptionally good. Trying to monitor year by year, subject by subject, on the basis of informal observations and snatches of rumour is particularly unsatisfactory. There is general agreement that year groups differ considerably so that simply looking for changes since last year is inadequate. Headteachers might manage to have a very good idea but it is the more remote inspectors/advisors who are assigned responsibility for monitoring. (Coopers and Lybrand, 1988; GB.DES, 1988.)

It is certainly arguable that there are no 'experts' whose expertise has been validated. Furthermore, recommendations from one set of experts are often at variance with the recommendations of another set. Where is the evidence for the reliability and validity of inspections? What is their impact? What is the cost? Would quantitative monitoring yield results consistent with results from inspection methods? These are complex issues which must be dealt with at length elsewhere (see, for example, Wilcox (1990)) but suffice it to say that there was sufficient concern to warrant an attempt at quantitative monitoring.

What can inspectors/advisors offer if their asserted expertise as judges of effectiveness is questioned or rejected? Since monitoring systems are inevitably limited, and will also need to develop and change over time to take account of changing circumstances, there will always be a role for on-the-ground, on-site, inspection. Since inspectors, in the course of site visits, can observe aspects of schooling which even the most comprehensive monitoring might overlook, the contributions of inspectors and monitoring systems can be expected to be complementary rather than interchangeable.

If asking an expert is not an alternative to monitoring, we need to agree some measures of what is happening in classrooms, departments and schools, and measures of effectiveness relating to the goals and purposes of the system.

There are two ways of viewing both a system of inspections and a monitoring system: as *accountability* or as *feedback*. Before considering the pros and cons of each, it is worth asking whether teachers actually need feedback regarding their effectiveness?' Do they not just know if they are doing a good job? There are two questions here: Do teachers need feedback about students' immediate responses to their teaching? Do they need feedback regarding their own effectiveness as teachers with regard to producing examination results?

As far as the immediate response of students is concerned there is immediate feedback of a kind, in the classroom. Thus by teaching, teachers learn classroom control and the management of classroom processes. The feedback with regard to these goals is immediate. However, it is very difficult for teachers to evaluate the cumulative effect of their teaching activities on the cognitive achievements of their students. A lone teacher has nothing to compare results against. Even when his or her students take external, blind-assessed examinations, such as the A-level examinations under consideration in this article, the results will depend heavily upon the students, the features of the examination that year and interactions between the two. The prior achievement, aptitudes and developed abilities of the students will obviously influence results, as will the severity of grading that year in that subject and the amount of differentiation in the examination. There may also be, quite accidentally, bias in the examination so that the items favour, say, boys, or students who have had certain experiences.

An individual teacher cannot make a proper allowance for these many factors when trying to evaluate the effects of his or her instruction. The feedback which is needed must compare the teachers' students with similar students in other schools or colleges, that year, on that particular examination.

It might be thought that in the matter of students' attitudes it was sufficient to have an anonymous survey of students and find what percentage were satisfied with their school experience. But what percentage would be a reasonable percentage to expect? And how does this vary for schools in different kinds of neighbourhoods, for colleges as opposed to schools, for schools in which parents pay as opposed to state-supported schools? One could simply choose a figure, set an arbitrary target, say 80 per cent, and regard anything else as a failure to achieve a reasonable goal. Yet it could well be argued that 80 per cent satisfaction was *unreasonable* if it was rarely achieved. Again, to interpret the data they would need to be seen in the context of comparable data from other schools and colleges. So, even with a scale which can be interpreted in a criterion-referenced fashion, there is a need for comparative data in order to provide teachers in departments and schools with feedback about the effects of their own systems of instruction and management. *Asking schools to set targets before they have adequate knowledge of current performance is like asking people to shoot at a target in the dark; a mischievous waste of school time.*. In short, feedback is needed in order to evaluate performance, and not just raw data but *interpretable* feedback, which generally requires data in the context of comparable data.

Does the knowledge base come into this discussion about the desirability or otherwise of having a monitoring system, and in particular about emphasizing accountability or feedback? Yes. There is an extensive literature on the effect of feedback on performance. Indeed, the Hawthorne studies, which led to the well-known term 'the Hawthorne effect', were reinterpreted as a demonstration *not* of the effect of giving attention but as the effect of *feedback*. In an article published in the prestigious journal *Science*, using a fine integration of qualitative and quantitative records from the original study, Parsons (1974) formulated and provided support for the hypothesis that the steadily increasing production was a result of feedback to the operators. Where feedback operated, people observed their own level of performance and found ways to improve it.

However, the knowledge base also contains, in social psychology, some warnings about monitoring as surveillance, monitoring solely for accountability. When monitoring is perceived as surveillance it may produce negative reactions (Deci *et al.*, 1982). Furthermore, if monitoring is linked with material incentives such as merit pay, there are hints that it could be destructive of intrinsic motivation (Lepper *et al.*, 1973).

Glass (1975) suggested that people may work best when warmly accepted rather than carefully monitored. Although he came down on the side of monitoring *programmes* carefully, if not *people*, the question he raised is not without some basis in many studies in social psychology. In particular, attribution theory led to studies on the undermining of intrinsic motivation by rewards and surveillance. A recent study (Butler, 1988) showed children's intrinsic motivation reduced by the giving of normative grades in feedback for tasks. Are there analogies to be drawn with performance monitoring? Is teaching intrinsically motivating? Do adults as well as children respond to some kinds of feedback with decreased rather than enhanced motivation? On the other hand, the reinterpretation of the Hawthorne effect (Parsons, 1974) is representative of a widely accepted concept of feedback on task performance as being motivating.

The work on intrinsic motivation in psychology may or may not apply with any strength, but it seems highly likely that feedback will have different effects depending on the type of feedback given. Eventually, the kind of generator–test procedure which Simon specifies should be applied to indicator systems, considering, for example, the effects of feedback of different kinds and varying intensities. This means, ideally, varying the costs, styles and content of information systems in the framework of an experimental design. Important variables would be:

- the balance between normative vs. informational feedback;
- the amount of process information;
- other aspects which have been shown to be related to the effectiveness of evaluation reports (Cousins and Leithwood, 1986): sophistication, decision relevance, consistency with users' beliefs, user involvement, relevance to problems and controversiality.

In summary, since the evidence from the knowledge base suggests that monitoring might turn out to be either beneficial and motivating, or destructive and demotivating, more research is needed. Is this just an example of the uselessness of social science, producing contradictory evidence? Is social science useless as a guide to action? Not at all. Again, it is a matter of design as opposed to science. The scientific knowledge base identifies opposing forces. The strength and effects of these forces will depend on other factors, particularly relating to how the monitoring is operated, presented, the climate into which it is introduced and so on. More research, of a basic-science variety, on these various factors would be helpful but just as engineers cannot rely on physics to predict what will work, in social science also, *the basic research will serve to point out important factors and raise issues, but the design activity, the creating and testing of actual systems will still be needed.*

This lack of a definitive knowledge base, the lack of 'laws' which will give unequivocal answers to questions such as 'Will monitoring be motivating or demotivating?' is not simply an indictment of the state of social science. The problem lies in the complexity of the systems whose behaviour we are trying to predict. A monitoring system can be implemented in so many different ways, with so many different effects, containing so many interactions between variables that the prediction of the final impact is impossible from first principles. The approach can only be to implement systems and to work on their design until they achieve the goals we want without the negative side-effects. This approach to the design of a social system is not unlike the problems that confront scientists once they move from the simple systems of classical physics into the complexities of a world in which chaos is as typical as order. Thus Simon writes about the design of a computer system:

> The main route to the development and improvement of time-sharing systems was to build them and see how they behaved. And this is what was done. They were built, modified, and improved in successive stages. Perhaps theory could have anticipated these experiments and made them unnecessary. In fact it didn't, and I don't know anyone intimately acquainted with these exceedingly complex systems who has very specific ideas as to how it might have done so. To understand them, the systems had to be constructed and their behaviour observed. (*ibid.*, p. 25)

So much for the crutch of theory. Computer scientists, working in infinitely simpler environments than education, have to proceed empirically, experimenting to study the performance of, for example, parallel systems; generating and testing, to see how they work. There is unpredictability everywhere, not only in social science.

In summary, I have argued that we live in a world of artefacts, that in education we are trying to achieve certain goals, driven by values. In the design of systems which will achieve these goals, we must utilize the knowledge base of social science and the design concepts which are beginning to be understood by researchers such as Simon. The notion that we shall soon have theories to guide actions may be a chimera.

The outer environment

In considering the outer environment in which an information system has to survive, we need to consider several contexts, each one the arena of certain stake-holders: the technical situation, the research *milieu*, the policy context, and professional concerns.

The technical context: computers and statistics

As management-information systems are developing in schools and as LEAs are moving into computer-based information systems, it becomes inevitable that large amounts of data will be stored and some of these data will be interpreted. The best defence against poor data are better data. Too often schools find league tables in local papers based on raw figures of exam passes. The best defence against such unfair comparisons is the existence of a database in which some allowance can be made for the differences in the intake characteristics of students in those schools. Computerization, statistical packages, information systems all make this kind of development in performance monitoring relatively straightforward, well within the bounds of the existing knowledge base.

The research context

The effects of schools on pupils is clearly central to the discipline of education. In 1966, Coleman implemented a national survey of schools conducted primarily with a concern for equal opportunities between children of different ethnic backgrounds (Coleman *et al.*, 1966). The major flaw in this study was in the lack of any credible outcome measure. The one used, the verbal ability of pupils, is not a variable likely to be sensitive to the impact of teaching. The conclusion that Coleman drew that schools made little difference was a widely discouraging conclusion for educators and should really have been seen simply as confirmation of what one might have expected: that schools make little difference to *ability* measures. The robustness of ability measures to instructional interventions is almost tautological. Speeded tests which require nimble thinking skills rather than knowledge are *ability* tests and should be little affected by instructional interventions. If, however, we can ask if the schools have an effect on the *achievement* of pupils, then the answers become rather different.

Given that the United Kingdom has an examination system which, arguably, measures the kind of authentic achievement behaviours that can be taught in schools, it is not surprising that the first major study that demonstrated possible effects of schools on pupils was in the United Kingdom. Rutter *et al.* (1979) studied 12 London secondary schools over 4 years. They looked at a broad range of variables, of inputs, processes and outputs.

The major outcome variables were achievement in externally set examinations, delinquency, behaviour and satisfaction. Two subsequent notable studies of considerable size were the Junior School Project (Mortimer *et al.*, 1988) and a study of multiethnic schools by Smith and Tomlinson (1989). Reynolds and Reid (1985) provide an overview of school effectiveness research in the early 1980s. School effectiveness research received a considerable boost with the creation of the International Congress on School Effectiveness and School Improvement with annual meetings held in London (1988), Rotterdam (1989), Jerusalem (1990), Cardiff (1991) and Vancouver (1992).

It was in this research climate of increasing interest in 'school effects' that a request was received from a school governor to comment on mathematics results for a school. There was little which could be said about a set of raw results from an unknown, unmeasured, set of students, yet the issue was clearly of great concern to the school. This prompted the initiation of what was envisaged as a small, part-time, personal (that is, unfunded) research study of school effects in 'the sixth form'. (The first 2 years of non-compulsory schooling are often called the sixth form.) Two LEAs were asked to give permission for an approach to be made for schools. It was made clear that any data collected and analysed would be confidential to schools. Permission granted, a letter to schools invited them to a meeting to discuss setting up a project to compare A-level results across schools, essentially to answer the question 'How good were our A-level results this year in each department?' Thirteen schools sent a representative to this meeting and 12 schools agreed to participate for what was projected to be a 6-year project looking at the two major subjects taken at A-level: English and mathematics.

Is should be noted that, right from the start, the notion of 'whole school' effects was rejected in favour of seeking effects at the level of the department, the unit which delivered instruction, the unit which might be seen as the smallest unit of management.

At this time the project was seen as a research project. It was also seen as a way of responding to genuine concerns of schools about their work and it was therefore planned that the data would be fed back into schools each year. The project was called Confidential Measurement-Based Self-Evaluation. The 'confidential' term in the title referred to the fact that schools chose code-names for themselves (such as Eton, Colditz) and all the output, summarized in data tables, was reported to all schools using these code-names.

'Measurement-based' referred to the explicit statement that no particular teaching style was thought to be effective, that the researcher claimed no special expertise on teaching style and simply wished *to measure* what related to effective outcomes. The neutrality of measurement emphasized that this was not a judgemental exercise but an exploratory investigation to provide feedback. Further assurance was offered in the title by the words 'self-evaluation'. It was clear then, and it is still clear, that researchers from outside the schools cannot know the detail of the events of a school. Therefore, *the measures of effectiveness have to be interpreted by the school in the context of what they know about their own operations.* Indeed, by focusing on the department as the unit of analysis right from the start, the project focused on the unit of management which delivers the results. The data do not relate to the effectiveness of an individual teacher because, for most A-level examinations, two or more teachers have been responsible for the teaching. It is the department which must manage the teaching resources and build on the strengths and avoid the weaknesses of every teacher in the department. Whilst it is true that the techniques of performance monitoring used in this project could be used to evaluate the work of an individual teacher, this use is not recommended. The work of each teacher is dependent on the work of the entire department and the school. Furthermore, the data are probably of insufficient accuracy to be asked to consider the work of a single teacher. Furthermore, outcome measures are not decisive. A very good set of results could conceivably be obtained if the teaching were perceived as so poor that students turned *en masse* to private tutors.

The zietgeist *and policy context*

With the collapse of the highly centralized, state-planned economies of Eastern Europe, the status of the free market received a boost. Of course, the fact that one highly centralized, authoritarian system, full of disinformation rather than open feedback (violating all Simon's design principles) was unviable and harmful does not by itself imply that existing alternatives are all goodness and light.

The following quotation from an editorial in the *Sunday Times* (9 September, 1990) indicates the climate in which schools are being asked to operate:

> No real progress can be made so long as the people who currently run our education system remain in charge. We need to sweep away the bureaucracies and teaching unions (both are part of the problem, not the solution) by devolving the decisions and the resources to the schools themselves, which would then be free to compete for pupils however they liked [*sic*]. Parents would make their choices, and that, in turn, would reinforce good schools and drive out the bad. The writing is on the wall for our schools; if we do not act now, there might soon by nobody able to read it.

Florid, off-the-cuff designs for systems have to be considered with caution. Can good and bad schools actually be located or are there only good and bad departments? It was noted above that an initial assumption was that effects should be considered at the level of the school *department*, not the whole school. The data have been found to be reasonably supportive of this choice: most schools contain both 'good' departments with positive effects and other departments which appear less effective.

The concept of the free market operating in education needs evaluation by people with a sound grasp of the knowledge base in economics. Here it will simply be noted that apparently the exponents of the free market — Adam Smith, Malthus, Ricardo, James Mill, McCulloch — 'were all believers in state involvement. These clearly saw education as a public good, in the provision of which both state and local authorities had a central role to play.' (Simon, 1990; p. 26) (not H.A. Simon, but Brian Simon, and see also (Brian) Simon (1991)).

In short, the great exponents of the market economy did not see education as an appropriate realm for the operation of this competitive model.

LOCAL MANAGEMENT OF SCHOOLS

Another feature of the political ethos during the development of the ALIS monitoring system was 'LMS', local management of schools, or site-based financial control for schools and colleges. The devolution of budgets and management decisions to individual school units is an issue quite separate from the operation of competition between schools. Keeping decisions on the allocation of resources within one level of the system close to those who deliver the output of that level may be an effective strategy. It seems consistent with Simon's notions of semi-independent components. The area needs research but also design, the design of systems of local management which allow schools to achieve goals with a minimum of harmful side-effects.

PERFORMANCE INDICATORS

It was a report on the Local Management of Schools commissioned by the Department of Education and Science from the accountancy firm Coopers and Lybrand which brought the term performance indicators to the forefront in education (Coopers and Lybrand, 1988). This linking of local management with performance indicators is important. The unstated implication is that by devolving responsibility and resources to the school level, schools are enabled to manage their affairs in order to reach goals and that the achievement of these goals can be monitored by the next level up in the hierarchy, this being the LEA. This was the model put forward in the Coopers and Lybrand report. *It was not a model of competition between schools but of monitoring and quality assurance across all schools.*

The Coopers and Lybrand report was followed by publications from other accountancy firms providing various long lists of performance indicators (PIs), veritable dogs' dinners. There seemed to be considerable confusion surrounding the meaning of the term performance indicator. Items of management information were referred to as performance indicators along with measures of outcomes. Some resolution of this terminology would seem to be needed.

Richards (1988) proposed a division into three types of indicators: performance monitoring, compliance monitoring and process monitoring. The distinction between performance and compliance monitoring seems to be particularly useful. Compliance monitoring answers the question: 'Are certain required processes being implemented?' Thus, if a National Curriculum must be implemented, a check on the extent to which it is implemented is a compliance monitoring exercise. If equal opportunities are of concern and schools must have explicit policies written down and implemented, then checking on these policies is a matter for compliance monitoring: checking on the *implementation* of certain policies.

The *effectiveness* of these policies in achieving certain outputs moves into the realm of performance monitoring or outcome monitoring. It is in that area that the ALIS operates, the monitoring of outcomes: that is, effectiveness. It would be generally helpful if PIs, compliance indicators and budgetary data were not confused.

UNIVERSITY ENTRY POLICIES

Since A-levels are designed as an examination pre-university to form the qualification base that permits entry to the university, university-entry policies are important in the system which is monitoring A-levels. It should be noted that during the years in which the project has been running the number of 18-year-olds has declined rapidly. Thus the pool in which universities recruit has shrunk. Another change in the outer environment is in the examinations taken at the age of 16. At the beginning of the study these were of two kinds: Ordinary level GCSE was taken by no more than the top 40 per cent and an easier exam, the Certificate of Secondary Education, was taken by most of the remaining students in a cohort. The year 1988 saw the entire country switch to the GCSE exam, a single examination across the ability range with various provisions in some subjects, such as mathematics, for entry at different levels of an examination. The GCSE exam was not only designed to test across the ability range but was also different in style. In some subjects some of the marks were awarded by teachers for course work and these were added into the examination mark awarded for written papers. Only in mathematics and French was there no element of course work involved in the examination in 1990. It was widely suspected that GCSE was easier than the previous examinations but it is perhaps more accurate to say that it was different. There is evidence that it favoured girls and there is a distinct possibility that it favoured middle-class students who would receive support at home for the course-work element. (Had adequate monitoring systems already been in place for the examinations taken at 16 years of age. we would have extensive evidence on these matters.)

Over the course of this study it was also anticipated that there would be changes to A-levels. There was a report from the Higginson Committee (1988) suggesting that A-level standards be maintained but that the amount of content be reduced to make the exam broader. There was also encouragement for students to do five rather than three subjects by using the Advanced Supplementary Examination, which was designed to count as half an A-level: it would be as difficult as an A-level but cover only half the content. The 'gold standard' of difficulty of A-level examinations was to be maintained (Hughes, 1989). Although rejected, the recommendations of the Higginson Committee would seem to be far from dead.

EXAMINATION BOARDS

Reference has been made to the fact that there is almost no good outcome measure available in the United States for the effects of cognitive instruction. The existence of scholastic aptitude tests (SATs) is no substitute for curriculum based examinations. 'SATs' in the United States were timed, paper-and-pencil tests, which, arguably, simply measure developed abilities through the medium of very general areas of knowledge. The inadequacy of such examinations is being increasingly recognized in the United States and there is a search for 'authentic, high stakes' testing (Shavelson, 1990). Moreover, the SATs themselves are changing.

The United Kingdom has had what can rightly be claimed as authentic, high-stakes testing for decades. Examination Boards publish syllabuses. Schools choose to teach to the syllabuses. Teachers are hired to work with professional examiners in setting the examinations based on the syllabuses. The examinations are administered under strict conditions at various examination centres, at the same time and on the same day in schools and colleges throughout the country. Examination scripts are returned to the examination boards where teams of teachers are hired to mark the papers and grades are assigned in a series of examiners' meetings in which substantial efforts are made to 'maintain standards' at certain key positions on the grading scale. It is a well-worked out system which has been used in the United Kingdom and around the world for many decades. There is room for improvement but the system has credibility and fairness to a degree not achieved by alternatives.

For A levels there was, for many years, a distribution advised by the Secondary Schools Examination Council. This strange distribution had as its mode a failing grade: 30 per cent of the most able candidates, who had chosen to stay on and were often taught in small classes by the best qualified teachers, were failed at the end of 2 years of hard work. This peculiar practice has come under increasing attack (Fitz-Gibbon, 1985; Howson, 1987).

The important point about the Examination Boards is the chance of equity in the assessment of students' work. Personal bias cannot enter into the grading process since teachers are not marking the work of students whom they know. One small but vital change is urgently needed, however: candidates names should be removed and replaced with numbers in order to remove any potential for ethnic, gender or other bias from the grading of examination scripts.

Another aspect of prime importance for any assessment is that what is measured is what has been taught. It is a game that all have agreed to play. Furthermore, the examination items are not wholly multiple choice. Students undertake authentic tasks such as writing essays on topics and working out mathematical problems. There are points available in the marking scheme for partial solution to problems not simply right or wrong answers. Furthermore, these are certainly high-stakes tests: a single grade on an A-level examination can mean the difference between getting the place at the university that the student has applied for and being rejected. Much university entrance policy works on summing up the points from A-level grades across the subjects taken by students. For example, entry to some subjects may require 12 points, the equivalent of 3 Bs. The setting of these entry requirements is left to each individual university department. Their offers have to be adjusted with a view to the number of places for which the state will pay the university fees. Throughout most of this project, university fees and 100 per cent subsistence allowances were paid to students. In the last year or two, loans have been made to top up the grant which students get when they are at university. The 'high stakes' then, apply to the student, who receives substantial support for 3-year degree courses, and to society which invests heavily on the basis of this selection process.

TECHNICAL AND VOCATIONAL EDUCATION INITIATIVE (TVEI)

A prevailing breeze in the winds of change has arisen from the TVEI project (see, for example, Hopkins, 1990). This centrally sponsored initiative aimed to change the curriculum balance towards vocational and technical subjects and also to change interaction patterns in schools towards more student-centred and active learning. The initiative influenced the choice of process variables that were included in the ALIS project. What was not foreseen was the extent to which the existence of TVEI funds would make support for ALIS possible. At the time ALIS started to grow it was seen by many teachers as a move into a more objective consideration of teaching practices. The initial enthusiasms of TVEI were giving way to requests for evidence. The feelings expressed by one TVEI coordinator must have been echoed among others: 'We've been saying "teach in this way" but we haven't really got any evidence about the effectiveness of these methods.' The ALIS project offered a way into these questions by providing feedback to each department, each year.

Furthermore, because TVEI had brought teachers, researchers and LEA personnel together in numerous workshops, informal links existed which facilitated the growth of the project (Huberman, 1990).

TIME PRESSURES

Another aspect of the 'outer environment' in which any indicator system would have to survive was that teachers were under enormous pressures of time. Time pressure is inherent in an open-ended, never-finished job like teaching but conditions were particularly difficult due to the introduction of the National Curriculum, the need for such activities as fund-raising, reading publications from the Schools Examinations and Assessment Council (Wragg, 1991), writing development plans and developing glossy brochures. *In this climate there was clearly a need for a monitoring system which was very little trouble and took up very little of anyone's time.*

The preceding paragraphs have indicated the outer environment in which the monitoring system for A levels was designed and developed. We turn now to the inner environment, the mechanisms of the artefact.

The inner environment: rationale and research methods

As we turn to the design of the monitoring system and how it functions, the inner environment of the artefact comes under consideration. A major concern in this section will be to show how one particular knowledge base, the traditional content of research methods courses, provides the techniques, the concepts, the constructs necessary for the design of a performance monitoring system. Drawing on these constructs provides a common language which is widely understood. *Educational researchers, or social scientists well trained in quantitative methods, are ideally placed to create monitoring systems.* The design of a performance monitoring system can be summarized under the same headings as a piece of research: rationale, sampling, unit of analysis, choice of variables, validity, reliability, reactivity, interpretation, dissemination and utilization.

The only element present in that list of headings which is rarely present in research reports is 'reactivity' — the impact of the research activity on behaviour. This familiar research concept becomes particularly important in the design of indicator systems and will be referred to as the behavioural impact. *The* behavioural impact *is the most important aspect of any indicator system.*

Rationale

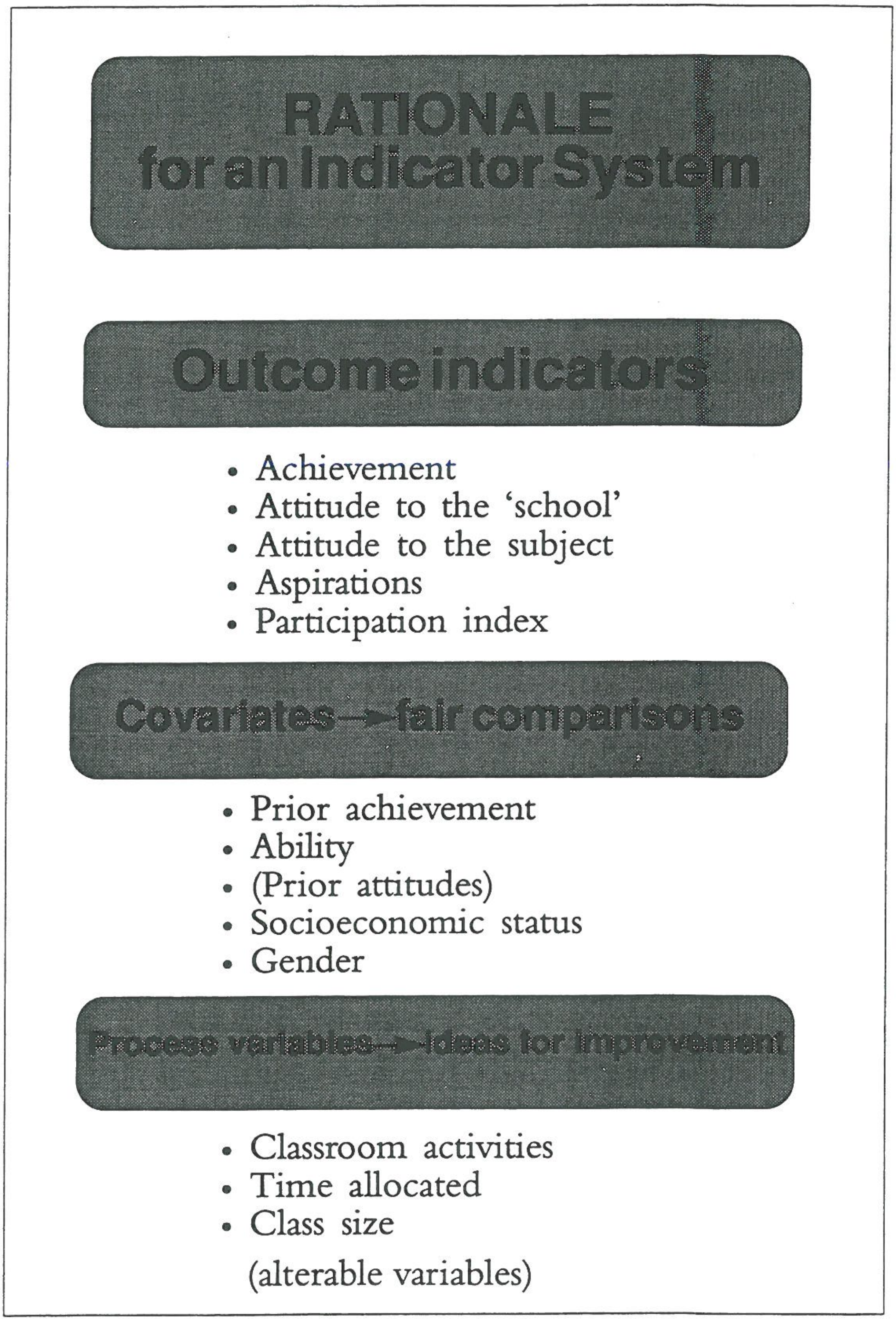
If the dog's dinner syndrome is to be avoided we need to set up an indicator system in which information is collected and interpreted in the framework of a rationale (see Fig. 2). This means that we do not take in just any data because they are available. To look at such data may be a waste of time. Conversely if needed data are not available, then a specific effort should be made to go out and collect them.

The rationale in which the ALIS project was formulated took as its starting point the goals or outcomes of concern. This is one of many junctures where value judgements come into the design of an indicator system. There must be agreement among those concerned, teachers and administrators, evaluators, local authority personnel, governors, etc., on the goals of the system. If there were no agreement then whoever had the most power would have to set the goals or would have to act to settle disputes in negotiations regarding goals.

In fact, agreement on goals does not seem to be an area of difficulty. The goals or outcomes agreed for the ALIS were:
achievement on the A-level examinations,

- attitude to the subject,
- attitude to the institution (school or college),
- aspirations for higher education,
- participation in extramural activities.

The last goal is a quality-of-life indicator that arises from the agreement that the last 2 years of school should not be simply 2 years of preparing for examinations. These years continue the personal development of the student and some indication of such personal development and quality of life can be obtained from examining participation rates in extramural activities, (such as orchestras, choirs, visits to the theatre, sports teams, chess clubs).



RATIONALE for an Indicator System

Outcome indicators

- Achievement
- Attitude to the 'school'
- Attitude to the subject
- Aspirations
- Participation index

Covariates → fair comparisons

- Prior achievement
- Ability
- (Prior attitudes)
- Socioeconomic status
- Gender

Process variables → ideas for improvement

- Classroom activities
- Time allocated
- Class size

(alterable variables)

Fig. 2: Rationale.

The rationale continues with the argument that if outcomes are to be examined, covariates need to be considered. Such covariates are those characteristics of students, or of the school context, over which the school has little control. Thus, if examination results are one of the goals, they should be considered in the context of prior achievement and developed abilities, variables which correlate highly with examination grades. The covariates considered in the ALIS project were prior achievement, developed abilities, gender and home background.

Once the outcomes have been considered in the context of the inputs, we have obtained measures of how effective departments were along a number of dimensions. This alone would constitute an indicator system. But the desire to provide feedback that is useful to practice led to the inclusion of process variables in the system, to see if any of the process variables were associated with effective cognitive instruction and/or positive student attitudes.

This third section, of process variables, could be infinitely large. The variables that have been considered in the ALIS project are class size, time allocated to instruction and various *alterable* classroom processes such as the frequency with which past exam papers are used for practice, the frequency of use of dictated notes, the frequency of assigning essays, the frequency of having students present their work to the class. (Some of these process variables were used in a study of school effects in Scotland (Gray *et al.*, 1983; p. 96).)

Sampling

Sampling of students is generally unacceptable to teachers in an indicator system. Teachers work closely with each student and want to see the results student by student. In the ALIS project the aim has always been 100 per cent response. Since there are inevitably some students absent on the day in which data are collected, questionnaires are left with freepost return envelopes. It appears that, as departments are in the ALIS project for longer, they become more concerned to ensure 100 per cent response from students. Fortunately the transparency of the system is important here and if, despite all efforts, a 100 per cent response rate is not obtained, departments can fill in missing examination indicators using the equations in the reports.

Unit of analysis

A decision was made at the very beginning of the project to look at results subject by subject, not to aggregate them. It seemed highly likely that the departments teaching each subject (for example, English, mathematics, economics) would vary in their effectiveness, and that to simply add up a variety of departments into some overall average would be to obscure information that was needed in managing the school. As described earlier, *the unit of analysis was the same as the unit of feedback and the unit of aggregation and these were consistent with the unit of management*, the school department. Pupil-level data have always been used and aggregated to the level of the school department in reporting back on effectiveness.

Not only would aggregation at the school level make the information less informative by hiding variations within the school, it would also promote the notion that there are overall good schools and overall bad schools. The data do not suggest that this is the case. Departments vary within schools more than schools vary among themselves and intercorrelations between indicators are low (Tymms 1990; Fitz-Gibbon, 1991).

The strategy used in developing the information system is to report on examinations, attitudes and processes every year to every department in every school or college that is participating.

Choice of variables

The constructs to be measured have been indicated in the rationale. Exactly how those constructs are operationalized is dealt with in this section on the choice of variables.

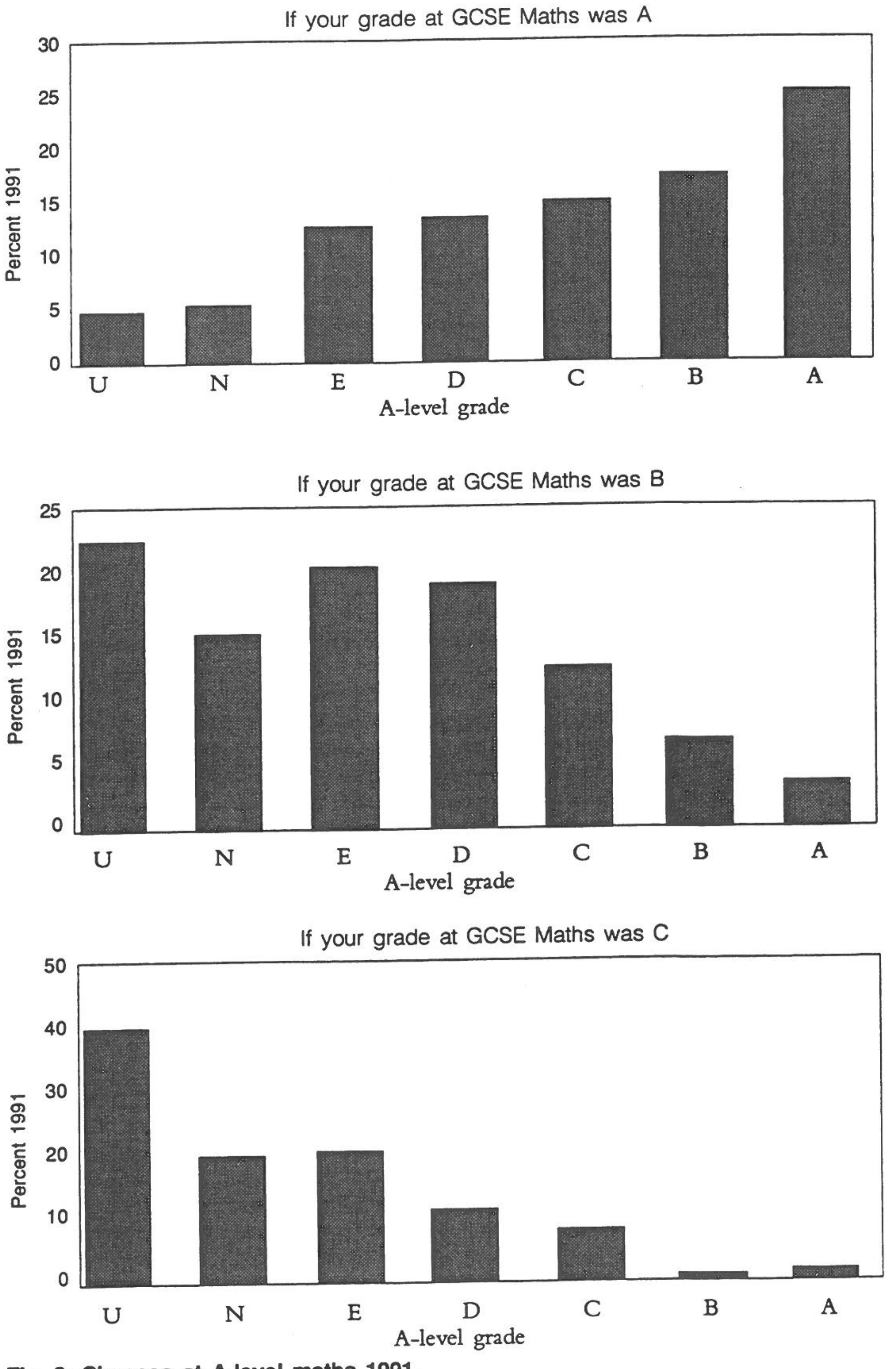
THE MEASURE OF PRIOR ACHIEVEMENT

The obvious measure of prior achievement was achievement in the external examinations at age 16, 2 years before A level. However, these examinations were set in a number of subjects. Some pupils may sit only one or two GCSEs, others may sit 10 or 12. How should these data be handled?

When consideration is given as to whether a student should proceed to an A-level course in a particular subject, the grade of that student in the same subject at age 16 is often looked at. Thus in the Cockcroft Report (1982) it was indicated that students who had not achieved a C in O-level mathematics were unlikely to pass A-level mathematics. In some schools or colleges students with a C would be debarred from A-level courses. They would either have to choose some subject other than mathematics or repeat their age-16 exam to attempt to get an A or a B. Figure 3 shows the proportions obtaining various grades in A-level mathematics in 1991 after a prior performance of A, B, or C. It can be seen that this situation has changed since the Cockcroft Report.

A C in GCSE, whilst not entirely promising for A level, is translated into a passing grade by almost half the candidates who are allowed to sit. Those allowed to sit for A level after obtaining a C will not, of course, be a random sample of those who attained a C. Other factors will have been taken into account by the student and the school. However, the data do call into question rules imposed by some institutions against students attempting A level without a prior A or B grade.

The prediction of an A level from the grade obtained in the same subject 2 years earlier is not, however, the best prediction possible. It is a general measurement principle that other things being equal the longer the test the more valid the estimate of the true score. One would expect, therefore, that a longer test would produce higher predictive validity. It was predictable from this principle that the average grade obtained in all the age-16 examinations would be the best predictor for the age-18 examinations. It could further be argued that performance at A level is multifaceted. No single subject taken at the age of 16 captures the broad range of skills required for A-level work. If instead of the single subject taken at age 16, one uses, for the prediction, the average across all the subjects taken, then the correlations are higher in almost every subject every year. There might be objections that students who only took one or two subjects at age 16 were at an advantage over those that took 10 or 11. However, those taking more subjects generally did better than those taking fewer subjects. Time is not the major determinant of achievement. Furthermore, the number of subjects taken is heavily influenced by school policies. In some schools it is not possible to take more than, say, seven exams at age 16. To avoid confounding school policies with what is meant to be a measure of general academic aptitude it is important to take the *average* grade obtained by the student at age 16 as the predictor, not a score based on the *total* grades obtained.



If your grade at GCSE Maths was A

If your grade at GCSE Maths was B

If your grade at GCSE Maths was C

Fig. 3: Chances at A-level maths 1991.

Of course the relationship needs looking at for non-linearity. This was done every year early in the project and deviations from linearity were not such as to merit the shift to a more complicated prediction equation. Changes in the second decimal point could be obtained by weighting A and B grades by, for example, taking a fourth power of the average score. Here, however, an important value judgement comes into play. *It is desirable that the monitoring system is as transparent as possible to the users.* By using a simple average grade as predictor the data become immediately accessible to those in schools who can handle equations such as $y = mx + c$. Indeed, teachers can and do compute residuals for their own pupils on the basis of the equations as provided in the report. Recently, to save teachers that effort, we have provided each school department with a printout of the average GCSE grades, the A-level grade and the residual for each pupil. This enables further analysis to be done at the school level and the influence of outliers is made apparent. (This development followed field-work in Cambridgeshire undertaken by David Elsom in connection with a PhD and his work as an adviser.)

THE MEASURE OF DEVELOPED ABILITY

Why was a measure of ability needed? Because in schools in which the students were first prepared for the age-16 exam and then for the age-18 exam, there was a very strong possibility that good results at age 16 would suggest that they were working with more able students than in fact they were, making it difficult to demonstrate a high level of 'value added' at age 18. A counterbalance to this problem was provided by the existence of ability tests administered to all students in the project under carefully standardized conditions. Tables of 'value added' using the ability test as a covariate are provided as an alternative to the tables based on prior achievement. The ability test chosen after trials of other tests was one kindly permitted to us by the International Association for Educational Assessment (Ottobre and Turnbull, 1987).

THE ATTITUDE SCALES

The measures of attitudes were developed from a large number of questions having face validity. Students selected responses on a 5-point scale from 'not true at all' to 'very true', to statements such as 'I like school', 'I would recommend this school or college to others'. The validity of the scale of attitude to school should be checked by in-depth interviewing but this has not yet been possible on a large scale. However, students' responses to an open-ended question have been compared with the quantitative scale and this has lent support to the scales (concurrent validity) (Fitz-Gibbon, 1989)

Validity

Mention has already been made of the construct validity of the examinations. Concern for the validity of the attitude measures has made it seem unwise to have schools themselves administer the questionnaire and the ability test. The credibility of the data is at stake here. If teachers are in the room, likely to look over the shoulders of students, or if teachers are likely to see the completed questionnaires, then students do not feel free to answer in an unguarded manner. From the second year of the project, therefore, the data have been collected by someone from the University. Small grants from the Department of Education and Science were very helpful at a critical juncture. They paid for the hire of data collectors, who were generally retired teachers or teachers undertaking supply work. Data collectors are trained, they are provided with a tape recording and detailed instructions and the materials that they need. They are paid *per diem* plus travelling expenses. This is a major expense in the project but it is felt to be essential for the credibility of the data. Furthermore, this method of data collection means that the project interferes very little with the process of the school. The demands on the staff time are an absolute minimum, amounting to no more than calling the students together in a room in which examination conditions can be obtained and leaving the data collector to administer the questionnaire and ability test. Returning the A-level results in August is the only other demand on the school's time. (There are also in-service training days available to help schools to interpret the data.)

Reliability

The reliability of examination data is the responsibility of Examination Boards and must be interpreted in the context of the reliability of feasible alternatives.

As for the reliability of the attitude measures, those items that did not correlate with other items in the scale were dropped in the early years of development. Since 1984 the scales used for attitude to the subject and attitude to the school have not changed. It is important in an indicator system to maintain the same measures from year to year for the purposes of comparability across the years. The reliabilities of the scales have remained reasonably high, with internal consistencies from about 0.6 to 0.8.

The reliability of the reports of classroom processes is entirely more problematical than the reliability of the other measures. There seems to be some disagreement within classes as to the frequency of use of various classroom processes. It should be noted that the frequency scale used is a fairly precise one. It is not simply 'never', 'rarely', 'sometimes', 'frequently', but rather:

- 'never or almost never'
- 'about once a term'
- 'about once a month'
- 'about once a fortnight'
- 'about once or twice a week'
- 'about every lesson'.

Obviously a student's recall of what has being going on in the class over the 2-year course is not precise and accurate. These process variables nevertheless correlated with some outcomes, suggesting that at least some perceived processes might relate to the outcomes, an hypothesis strengthened when the correlations were replicated in more than 1 year's data. The interpretation of these relationships would need a large-scale research project and experimental research. They are treated in the ALIS reports as interesting hypotheses, a basis for discussions among members of staff. For example, for the 3 years in which we have been analysing data on A-level chemistry (1988, 1989 and 1990) it has been found that students reporting frequent use of essays in chemistry tended to have higher residuals than those reporting less frequent use of essays in chemistry. This finding has been briefly discussed elsewhere in terms of generative learning, gender effects, student-centred learning and preferred approaches to science (Fitz-Gibbon, 1991). It is important to stress that correlational data do not establish causal relationships but the value of this finding lies in the discussions generated among chemistry teachers as to what students have understood by the term 'essay', how teachers set, mark and use essays and the structure of the A-level chemistry examination. Furthermore, inspired by the quantitative hypothesis, teachers can try using more essays and observing the impact on not only on learning but also on student attitudes, especially since more positive attitudes were found in chemistry classes reporting frequent use of essays.

The way to establish both the validity and the reliability of the process data is to use them as a basis for interventions. If altering a process variable can demonstrably change the outcomes, this establishes an important finding as well as the validity and reliability of the measures that suggested the intervention. Work of this kind is in progress (Tymms and Fitz-Gibbon, 1991).

Reactivity: behavioural impact

The problem of the influence of measurement on that which is measured has long been recognized in physics and is well recognized in social science where it is termed 'reactivity' (Campbell and Stanley, 1966). The impact of measurement on the A-level students is unlikely to be particularly strong. They are only made aware of this monitoring process on one occasion, when they complete the questionnaire and take the aptitude test. The effects of the monitoring are greatest on the staff of schools and colleges participating. Indeed, the feedback is meant to produce reactions in school management teams, reactions aimed at maintaining quality in every department. The possibility of demotivating surveillance as opposed to motivating feedback has already been discussed, and it is emphasized that monitoring must, crucially, produce positive behavioural reactions.

There is, of course, widespread agreement that monitoring should be beneficial in its impact. What is not so well recognized is the effect of the choice of variables on this impact. For example, if percentage pass rates are used as indicators, then the logical response is to teach towards the borderline group of students, for it would only be there that the indicator could be improved. If a student thought to have the potential of an A got a C, this would not affect the percentage pass rate so such a student might as well be largely ignored. If, on the other hand, a borderline candidate scraped through, the indicator would improve. Thus the behavioural implications of percentage pass rates are to concentrate teaching on borderline candidates. This is educationally poor and unjust. *Each pupil should count equally in the production of PIs, so that equal importance is given to each pupil.*

Interpretation

The design used in the performance monitoring system is clearly a correlational one, a repeated survey, and this places some severe limitations on the interpretations that can be drawn (Wainer, 1989). 'Correlation is not causation' should haunt the consciousness and the conscience of everyone trained in research methods (cf. Preece, 1989). The tentativeness of the relationships between examination residuals, attitude data and classroom process data is constantly emphasized.

Dissemination and utilization

The most comprehensive database in the world will be of little use if it is uninterpretable, and even a database developed to a rationale that leads to clear indicators will be useless if it is not interpreted for those who need to understand it.

In the ALIS project, workshops are held to begin to develop, in each school or college, a number of staff members who are fully conversant with the system.

Drawing on Simon's theories of effective organizational structures, on the Educational Reform Act legislation and on the Coopers and Lybrand report, it would seem clear that the communication and interpretation of the results should eventually become the responsibility of LEA inspectors and advisers. Indeed, since the project was renamed ALIS in 1988 it has been an explicit, stated, written aim of the project to transfer expertise into the LEA. The expertise which is most urgently in need of transferring is the capacity to interpret the data. Eventually, some LEAs may take over the data analysis procedures if they have sufficiently skilled research and statistics' branches. Examination boards could provide the value-added residuals and, indeed, in Scotland that will soon be happening.

Propositions/conclusions

There are two kinds of conclusion from the ALIS project. One set of conclusions is grounded in the data; in large part these detailed conclusions are being reported directly to schools and in research papers which have been published or are under development.

Other conclusions are of an experiential kind: that which researchers believe they have learned from the experience of ALIS. These are more in the form of propositions rather than conclusions and they test the ideas of H.A. Simon against some current concerns in education.

The need for experiments has already been mentioned. There are many experiments undertaken for research purposes, primarily in psychology, but not the naturalistic, field experiments which are needed for the generator–test cycles of the design process. Just as engines have to be tested under the conditions in which they will operate, so educational processes need testing under the relevant conditions. This calls for naturalistic, field experiments, as in clinical trials in medicine.

Whilst some work is under way to look at the effect of classroom processes which are very little used and yet seem to be related to good exam results (Tymms and Fitz-Gibbon, 1991), it is nevertheless recognized that intervention studies, important as they are, will take a long time to produce definitive answers, even assuming that definitive answers prove possible. The possibility of interactions between type of teacher and type of student and subject area and subject topic, all the complexities of the educational situation, cannot be underestimated. The need is for constant monitoring to provide practitioners, rather than researchers, with the feedback they need.

If this proposition of complexity is correct, what are the implications for UDEs and LEAs?

The role of UDEs (University departments of education)

As many UDEs celebrate their centenary, the problem of their identity continues to cause concern. That this identity crisis has become endemic in the discipline, and is not a problem confined to a few people in the United Kingdom, is indicated by the clear recognition of this identity crisis throughout the United States. Thus Clifford and Guthrie

(1988, p. 3) argue that schools of education '... have seldom succeeded in satisfying the scholarly norms of their campus letters and science colleagues, and they are simultaneously estranged from their practising professional peers'. In another volume, celebrating centenaries in the United Kingdom, Edwards notes the demise of the 'four disciplines' approach to education, in which education students were to be inducted into the disciplines of psychology, sociology, history and philosophy: 'The notion of a body of rational knowledge about teaching which students could first acquire and then apply in particular classrooms has been largely abandoned in favour of "reflective practice"' (Edwards, in Thomas (1990), p. 187).

H.A. Simon recognized this identity crisis in professional schools. He argued that it arises from the fact that the artificial has to be dealt with not directly by the methods of science but by the methods of design. Until recently there was no intellectually defensible methodology for design, so that people in professional schools, having started out by teaching their craft, such as education, engineering or architecture, then tried to attain higher status by being scientists rather than designers. This alienated them from the practice of their profession and from practitioners. 'The older kind of professional school did not know how to educate for professional design at an intellectual level appropriate to a University; a newer kind of school has nearly abdicated responsibility for training in the core professional skill' (H.A. Simon *op. cit.*, p. 131).

Many staff in departments of education will recognize the identity crisis. In the United Kingdom, criticisms that those training teachers were out of touch with teaching have led to legislation requiring them to spend one term in schools every 3 years, obtaining 'recent and relevant' experience.

The concept of *design* as distinct from *science* helps to resolve this identity crisis. *Our major task is to design systems that work.*

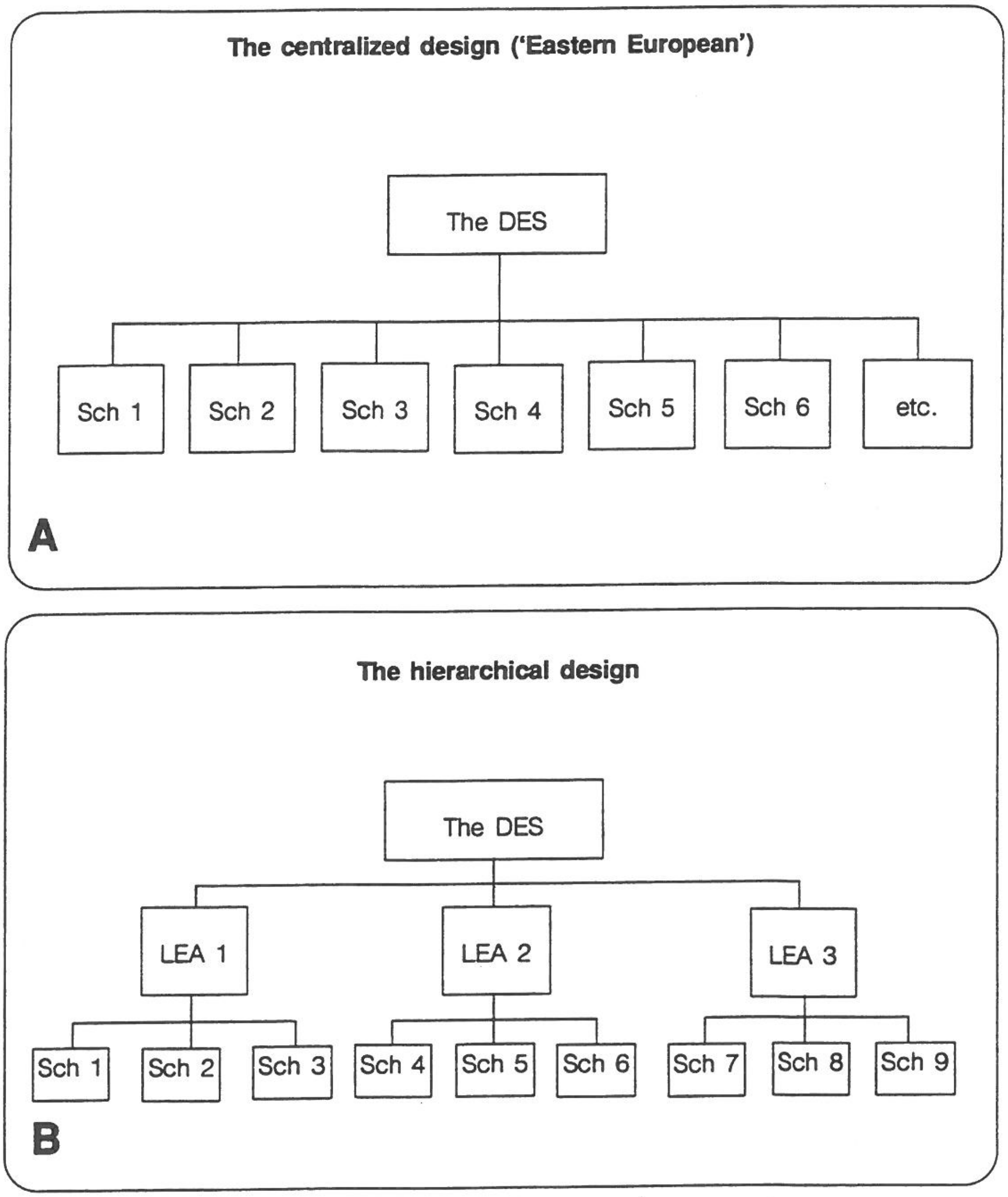
We must design systems for every module, every semi-autonomous unit of education: not only monitoring systems for schools and departments but also instructional systems for classrooms, and organizational systems for the country as a whole. If schools of education contribute to intellectually defensible *designs* for education, resting on more than 'everyday experience', then they justify their existence as professional schools. Their staff are not sociologists *manqué*, would-be psychologists, errant philosophers or limited historians but educational designers, social engineers. This conceptualization fits in with the reflective practitioner concept noted by Edwards and with Clifford and Guthrie's call for 'field-generated enquiries'.

Edwards (1990, p. 187) points out:

> Reflection is not a matter of recollecting in the tranquility of the university department the emotions, failures and successes of 'real' lessons; it involves disciplined inquiry into how learning is organized and assessed, and above all how the teacher can make his or her own methods an object for investigation.

Does focusing on 'reflective practice' solve the identity crisis? What follows from reflection? Not, surely, research-type conclusions, contributions to a formal body of knowledge, but, rather, revisions of practice, revisions in the details of the design of instruction, new crafting of learning experiences, the reconstruction of systems that achieve goals. If this *design* activity is *informed* as well as intuitive, if there is something to be learned about the design of systems that work, then there is a reason for the existence of schools of education, and the theoretical and the practical come together.

Just as an engineer uses laws established in physics, and techniques and instruments developed in 'pure' research, in the design of systems that work in the physical world, educational designers (whether called teachers or called researchers) need to build on the



The centralized design ('Eastern European')

The hierarchical design

Fig. 4: Two designs for education.

knowledge basis, such as it is, in psychology and to have scientific tools available such as statistics, measurement, design to help them to craft systems that work. Monitoring provides a scaffolding for this activity.

LEAs and monitoring: let's avoid an Eastern European model

University educationalists are not the only educational professionals being subjected to legislative attacks. Staff in LEAs, known variously as advisers or inspectors, are finding their jobs redefined and their budgets drastically reduced.

The notion that LEAs might wither on the vine causes concern if we take seriously Simon's concepts of complex systems needing to be constructed in a modular fashion, of semi-independent units and hierarchical structures. The design for education should not become like Fig. 4A, an Eastern European (before the changes) model, but rather should remain like Fig. 4B.

LEA personnel are needed to help with the interpretation of the data; to check yearly on the technical quality of the data collection, storage, retrieval, analysis and interpretation; to investigate the validity of the measures at frequent intervals — all this in addition to keeping a watching brief on the indicators and ensuring that deviations from high quality are remarked, and deficiencies remedied.

The approach which might be most effective in running a monitoring system is the approach of a humble scientist, a collaborative investigator with teachers. The LEA inspector should be asking: 'What seems to be related to effectiveness? What seems to produce good student attitudes? How can the alterable variables of the classroom be adjusted to achieve the goals of the system and to maintain quality across the system from year to year in every department?' This collaborative mode has no room in it for advisers and inspectors who claim special knowledge of 'good practice'. As already remarked, they may know what practice they like but until this claimed knowledge of good practice is substantiated by evidence it must be treated simply as an hypothesis rather than as established knowledge.

However (to back away a little from this over-harsh stricture), this is not to say that there is no room for values, for asserting that some processes are valuable in and of themselves, not because they are effective in any immediately measurable way. But a process which is valued intrinsically, rather than as a means to an end, should be the topic of *compliance monitoring* not performance monitoring (Richards, 1988). If students should have the experience of working in teams because so much work later in life depends upon working in teams, then working-in-teams should be declared desirable and checked for compliance. Arguments about the effectiveness of working-in-teams, whilst still important, may be temporarily shelved if there is wide agreement that this is an experience students need.

The question of teaching processes is, indeed, an area of great difficulty. It may be argued that certain styles of education, such as a highly practical mode in science, will yield results not at A level but in the subsequent years at university or when the student is out in industry and needs to operate in a team and in a problem-solving mode. It is certainly defensible to argue that an intelligent animal like a human being learns from experience and if a human needs to work in teams, then the experience of working in teams is precisely the experience that is needed to improve team work.

Whatever the issues, teachers should be recognized as co-equal investigators in the process of maintaining quality and looking for areas in which the whole process can be redesigned for improvement. Only well-established research findings should be promoted and these are few and far between.

Table 1: Dictated notes

Subject	Year	% students reporting 'frequent' use[1]	Reported frequency of use associated with:			
			Most positive attitudes	*p*	Most positive residuals	*p*
Mathematics	1988	77	Any[2]	0.68	Frequently	0.08
	1989	68	Occasionally	0.48	Occasionally/ frequently	0.21
	1990	71	Occasionally	0.55	Frequently	0.14
Geography	1988	83	Never	0.38	Any	0.99
	1989	88	Never	0.22	Frequently	0.18
	1990	81	Never	0.06	Never	0.60
French	1988	13	Never	0.59	Frequently	0.93
	1989	23	Frequently	0.61	Occasionally/ frequently	0.71
	1990	20	Any	0.82	Never	0.19
Chemistry	1988	83	Never	0.16	Not occasionally[3]	0.01
	1989	79	Never/ occasionally	0/02	Occasionally/ frequently	0.92
	1990	79	Never/frequently	0.70	Never/ occasionally	0.06

[1] The following scales were used: 'occasionally' represents up to once a month; 'frequently' represents up to every lesson; 'never' represents 'never' or 'almost never' reports.

[2] 'Any', no differences among the outcomes.

[3] There was a strong negative residual for the group which reported occasional use.

To take just one example, teachers are sometimes criticized for dictating notes. Is there justification for such criticism, for questioning teachers' strategies? In Table 1, the association between students' attitudes to the subject they were taking at A level and the reported frequency of use of dictated notes is shown for four subjects. (The four subjects are those used in Fitz-Gibbon (1991) and represent a language, a humanities subject, mathematics and a science). The reported use showed consistent levels across 3 years, that is, from three cohorts of students, with only French having 'frequent' use reported by small percentages of students (about 20 per cent). In the other subjects 70 per cent and more generally reported use of dictated notes more frequently than once a fortnight.

For each subject a one-way analysis of variance was used to see if there were differences in attitudes associated with different frequencies of use of dictated notes. Recorded in the table is the level of frequency which was associated with the most positive attitudes to the subject. There were only two statistically significant relationships at the 0.05 level. These did point to better attitudes associated with using dictated notes never or only occasionally (in geography and chemistry). Predominantly, however, the 'effects' each year were far from interesting levels of statistical significance. The table also shows the relationship between the use of dictated notes and students' residuals (positive residuals implied the students did better than would have been expected on the basis of their prior achievement.) Eight of the twelve sets of data could be taken as supporting the frequent use of dictated notes, although the individual findings were largely not statistically significant.

These findings should simply be taken as suggesting hypotheses. Only experimental approaches can establish causal relationships. However, the lack of strong support for a 'thou shalt not use dictated notes' *diktat* can be used to free teachers from being bullied by those who grasp at straws in trying to advise others on how to undertake their professional activities.

It might be argued that a few 'thou shalts' or a few 'thou shalt nots' do not do a lot of harm. For example, since Rutter *et al.* (1979) found that the display of students' work on the walls was associated with better achievement, some inspectors/advisers have been suggesting to teachers that it would be a good idea to put students' work on the wall. Does this do any harm? Perhaps not, or perhaps it does. Quite apart from the fact that inspectors/advisers are highly paid and such advice is not worth the cost, the main danger is that of treating correlational findings as though they were indicators of causal relationships: such unjustified interpretations delay the kind of research that needs to be done, just as medical quackery can delay the search for effective treatment.

The interpretation of correlational findings as causal led, in the United States, to the worsening of the already stressful working conditions of thousands of teachers. Drawing on the correlational, survey work of Wiley and Harnischfeger (1974) legislation was introduced to lengthen the school day and increase the number of days in the school year, this despite strong challenges to the validity of the findings (see Fitz- Gibbon (1989) for further details). Researchers who allow correlational findings to be interpreted as causal should at least be sure their poor practice is not harmful to others. The distinction between correlational and causal findings was emphasized no less than 14 times by Rutter and his colleagues. For example: 'Firm conclusions on causative influences can only come from experimental interventions' (Rutter *et al.*, 1979).

With the existence of an information system there is created a new and more intellectually defensible role for inspectors and advisers: the role of a collaborative designer working with teachers on the basis of a database which they and the teachers can renegotiate and alter to meet their information needs. It may be a far more comfortable role than attempting to impose procedures for which the knowledge base is insubstantial.

Action research: working with live data

The improbability of discovering useful laws in social science has been suggested above. As a further heresy let it be said that working on a monitoring system rather puts one off any return to working on dead data, data on which no action is to be taken. For example, finding gender differences in achievement or slight differences in the mean residuals for different types of institution (for example, comprehensive school, sixth-form college) become poor substitutes for providing information *to those concerned* about the gender differences, about the effectiveness — so that the live data can affect live decisions.

'But is it research?' An exceptionally competent ALIS research associate was asked this at a job interview. It has been argued here that *our major task in the professional schools such as education, medicine and engineering is to design systems that work.* And design is not research. The anecdote illustrates that this concept of design as our major professional responsibility is not a mere academic distinction. A conceptual shift is needed because without it resources will not be allocated and jobs will not be won for the task of improving education. In a sense, however, design projects turn out to be research projects: research into how systems behave.

Finally, a major conclusion from the ALIS project was described by Burstein (1991) as 'an existence theorem'. It was not necessarily obvious that a project would survive (for 10 years and very likely more) which measured the effectiveness of school departments and published league tables (albeit using code-names); which asked students about what went on in classrooms and their attitudes to teachers, their lessons and the schools or colleges attended; which also produced league tables on these aspects. The existence and continued growth of this information system proves something, perhaps, about the extent of professional and rational interest among teaching staff.

Notes

[1] 'A levels' refer to externally set and marked examinations taken, generally, at the age of 18 after 2 years of non-compulsory advanced study. These examinations are the primary hurdle for entry into higher education in the United Kingdom.

[2] Her Majesty's Inspectors (HMI) are employed by the Department of Education and Science to visit schools and make reports. LEAs also operate an inspectorate. The hallmark of the work of inspectors, local and national, has been the site visit: they spend several days in a school which they are inspecting.

[3] GCSE: General Certificate of Education. Externally set and marked examinations for pupils who are approximately 16 years of age, at the end of compulsory schooling. As with the A levels, these examinations are based on published syllabuses and could be said to represent what is discussed in the United States as authentic, high-stakes testing.

References

AITKIN, M. and LONGFORD, N. (1986). 'Statistical modelling issues in school effectiveness studies', *Journal of the Royal Statistical Society (Series A)*, 149, 1, 1–43.

BURSTEIN, L. (1991). Personal communication.

BUTLER, R. (1988). 'Enhancing and undermining intrinsic motivation: the effects of task-involving and ego-involving evaluation on interest and performance', *British Journal of Educational Psychology*, 58, 1–14.

CAMPBELL, D.T. and STANLEY, J.C. (1966). *Experimental and Quasi-experimental Designs for Research.* Chicago: Rand McNally.

CLIFFORD, G.J. and GUTHRIE, J.W. (1988). *Ed School: a Brief for Professional Education.* Chicago: University of Chicago Press.

COCKCROFT, W.H. (1982). *Mathematics Counts: Report of the Committee of Inquiry into the Teaching of Mathematics in Schools.* London: HMSO.

COLEMAN, J.S., CAMPBELL, E.Q., HOBSON, C.J., McPARTLAND, J., MODD, A.M., WEINFIELD, F.D. and YORK, R.L. (1966). *Equality of Educational Opportunity.* Washington: US Government Printing Office.

COOPERS & LYBRAND (Accountants) (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, 331–64.

DECI, E.L., SPIEGEL, N.H., RYAN, R.M., DOESTNER, R. and KAUFMAN, M. (1982). 'Effects of performance standards on teaching styles: behavior of controlling teachers', *Journal of Educational Psychology*, 74, 6, 852–9.

EDWARDS, A.D. (1990). 'Schools of education: their work and their future'. In: THOMAS, J.B. (Ed) *British Universities and Teacher Education*. London: Falmer.

FITZ-GIBBON, C.T. (1985). 'A-level results in comprehensive schools: the Combse project, year 1', *Oxford Review of Education*, 11, 1, 43–58.

FITZ-GIBBON, C.T. (1985). 'TERSE reports: towards experimental research synthesis in education', *Evaluation and Research in Education*, 2, 3, 147–8.

FITZ-GIBBON, C.T. (1989). 'Performance indicators: educational considerations'. In: LEVACIC, R. (Ed) *Financial Management in Education*. Milton Keynes: Open University Press.

FITZ-GIBBON, C.T. (1991). 'Multilevel modelling in an indicator system'. In: RAUDENBUSH, S.W. and WILLMS, J.D. (Ed) *Schools, Pupils and Classrooms: International Studies of Schooling from a Multilevel Perspective*, pp. 45–61. London and New York: Academic Press.

GLASS, G.V. (1975). 'A paradox about the excellence of schools and the people in them', *Educational Researcher*, 4, 3, 9–13.

GOLDSTEIN, H. (1987). *Multi-level Models in Educational and Social Research*. London: Griffin.

GRAY, J., McPHERSON, A.F. and RAFFE, D. (1983). *Reconstructions of Secondary Education: Theory Myth and Practice Since the War*. London: Routledge and Kegan Paul.

GREAT BRITAIN. DEPARTMENT OF EDUCATION AND SCIENCE (1988). Performance Indicators for Secondary Schools: Some Practical Considerations. A discussion paper presented at 'Quality in Schools: a briefing conference', NFER, 29 June, 1988.

HIGGINSON, G. (Chair) (1988). *Advancing A-Levels*. London: HMSO.

HOPKINS, D. (1990). *TVEI At the Change of Life*. Clevedon: Multilingual Matters.

HOWSON, G. (1987). 'A-level mathematics: some thoughts. (The opening address)'. In: EVERTON, T. (Ed) *The Reform of A-level Mathematics. A Report of the Conference held at the University of Leicester.* SMP 16–19 project, Leicester: University, School of Education.

HUBERMAN, M. (1990). 'Linkage between researchers and practitioners: a qualitative study', *American Educational Research Journal*, 27, 2, 363–91.

HUGHES, J. (1989). *AS Levels: Implications for Schools, Examining boards and Universities*. London: Falmer.

LEPPER, M.R., GREENE, D. and NISBETT, R. (1973). 'Undermining children's intrinsic interest with extrinsic rewards: a test of the "overjustification" hypothesis', *Journal of Personality and Social Psychology*, 28, 129–37.

MORTIMER, P., SAMMONS, P., STOLL, L., LEWIS, D., and ECOB, R. (1988). *School Matters*. Wells: Open Books.

OTTOBRE, F.M. and TURNBULL, W.W. (1987). The International Test of Developed Abilities: A Report on the Feasibility Study. Report for the International Association for Educational Assessment, Princeton, New Jersey 08541.

PARSONS, H.M. (1974). 'What happened at Hawthorne?', *Science*, 183, 922–32.

PREECE, P. (1989). 'Pitfalls in research on school and teacher effectiveness', *Research Papers in Education*, 4, 3, 47–69.

RAUDENBUSH, S. and BRYK, A.S. (1986). 'A hierarchical model for studying school effects', *Sociology of Education*, 59, 1–17.

REYNOLDS, D. and REID, K. (1985). 'The second stage: towards a reconceptualization of theory and methodology in school effectiveness research'. In: REYNOLDS, D. (Ed) *Studying School Effectiveness.* London: Falmer.

RICHARDS, C.E. (1988). 'A typology of educational monitoring systems', *Educational Evaluation and Policy Analysis*, 10, 2, 106–16.

RUTTER, M., MAUGHAN, B., MORTIMORE, P. and OUSTON, J. (1979). *Fifteen Thousand Hours: Secondary Schools and Their Effects on Children*. London: Open Books.

SHAVELSON, R.G. (1990). 'Can indicator systems improve the effectiveness of mathematics and science education? The case of the US', *Evaluation and Research in Education*, 4, 2, 51–60.

SIMON, B. (1990). The Future in Education: Which Way? Centenary Lecture, School of Education, University of Newcastle Upon Tyne.

SIMON, B. (1991). *Education and the Social Order, 1940–1990*. London: Lawrence and Wishart.

SIMON, H.A. (1988). *The Sciences of the Artificial* (2nd edn). Cambridge, MA: MIT Press.

SMITH, D.J. and TOMLINSON, S. (1989). *The School Effects: A Study of Multi-racial Comprehensives.* London: Policy Studies Institute.

THOMAS, J.B. (Ed) (1990). *British Universities and Teacher Education: A Century of Change*. London: Falmer.

TYMMS, P.B. (1990). The Stability of School Effectiveness Indicators. Paper given at the annual meeting of the British Educational Research Association, Roehampton, 1990.

TYMMS, P.B. (1991). 'Can indicator systems improve the effectiveness of science and mathematics education?', *Evaluation and Research in Education*, 4, 2, 61–73.

TYMMS, P.B. and FITZ-GIBBON, C.T. (1991). *Students at the Front*. Newcastle upon Tyne: The Curriculum, Evaluation and Management Centre, School of Education, University of Newcastle upon Tyne.

WAINER, H. (1989). 'Eelworms, bulletholes, and Geraldine Ferraro: some problems with statistical adjustment and some solutions', *Journal of Educational Statistics*, 14, 2, 121–40.

WILCOX, B. (1990). 'Is there a role for site visits in monitoring systems? A UK perspective', *Evaluation and Research in Education*, 4, 2, 81–90.

WILEY, D.E. and HARNISCHFEGER, A. (1974). 'Explosion of a myth: quantity of schooling and exposure to instruction, major educational vehicles', *Educational Researcher* (April), 7–12.

WRAGG, T. (1991). 'The relapse, ... or carry on testing', *The Times Educational Supplement*, 22 February, 104.

Correspondence

C.T. Fitz-Gibbon, School of Education, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU.