Curriculum, Evaluation and Management Centre

PUBLICATION NO.54

MONITORING WITH FEEDBACK: THE DEMOCRATISATION OF DATA

C.T. FITZ-GIBBON

1993

Comments on a 'Comparison of methods currently being used to evaluate schools in Scotland' by Dr. L. Paterson

MONITORING WITH FEEDBACK: THE DEMOCRATISATION OF DATA

by

Carol Taylor Fitz-Gibbon

School of Education, University of Newcastle upon Tyne

I find myself in strong disagreement with Dr. Paterson's main conclusion which is that data on school performance should be reserved for research purposes. Empirical evidence of the effects of various ways of using such data is yet to be gathered but there are many reasons to welcome the availability of performance indicators. Dr. Paterson's arguments seem to rest on an assumption about one style of management and on exaggerated reservations about the statistical adequacy of the data.

Before examining the reasons for rejecting these arguments let me emphasise that there are very wide areas of agreement between us. (Indeed, were there many more discussions and revisions to these papers the interesting cognitive conflict might be lost altogether! We both hope that the disagreements aired here are stimulating.) I share with Dr. Paterson an opposition to unfairly manipulative, macho-style management which is an ever-present danger ("the price of freedom is eternal vigilance") and I agree with all of the criticisms of raw data and "adjusted means" (also known as "means on means" analyses). The use of highly aggregated lumps of information to represent a system which can only be adequately modelled using fine-grained, student-level data was warned against in a now famous meeting of the Royal Statistical Society (Aitkin and Longford, 1986).

I agree entirely that teachers cannot be held responsible for outcomes over which they have no influence. Indeed I offered the slogan "no accountability without causality" at one of CES's esteemed international conferences (1988). But finding out what can and cannot be changed in a system requires information. Performance Indicators are an important type of information from which we can all, teachers, statisticians and researchers, learn together ...and perhaps even politicians can begin to learn the discipline imposed by data.

Synopsis of the arguments to be presented.

The remainder of this paper is sub-headed with the phrases highlighted in the following brief outline of the arguments to be put forward: Other **management styles** than the "domineering" one are possible and, even if one is saddled with a domineering style of management , it is better to have data than not. Dr. Paterson's treatment of **statistical issues** makes the odd and unjustified assumption that the data are to be used to classify schools and argues that the data are inadequate for this straw-man purpose. Moreover his argument implies that there is a critical type I error rate known to statisticians despite the novelty of this kind of data. This is unjustifiable. Indeed it is a fine example of "the great hubris of statistical research" against which Dr. Paterson warns us. **Relative ratings** are criticised for not being what no-one has claimed they are and their usefulness is not fairly acknowledged. They provide, for example, important confirmation that there are major differences between *departments* within schools rather than simply "good schools" and "bad schools". This evidence can undermine the political thrust for competition between schools and is just one way in which good data *does* "insinuate" **accountability** into national policy making in a way Dr. Paterson would wish to see happen. The data Dr. Paterson presents on Relative Ratings may also contain evidence consistent with increasing discrepancies among schools and could present a possible test of national policy.

The important issues which now face us in Education are about **how to run indicator systems** in ways which improve education, rather than about whether or not to have indicators. It is too late to bolt the stable door even if we wanted to.

Comments in more detail

Management Styles. Since much of his argument rests on the notion of the use of indicators for "accountability" it is unfortunate that there is no discussion of how Dr. Paterson imagines "accountability" will work. What does the term imply? Dr. Paterson seems to operationalize it as a process of classifying schools, but this straw-man is set up without justification. More on the statistical aspects of this later. For now let it be suggested that the existence of poor styles of management is not due to the existence of performance indicators or numerical data. Indeed, domineering, carrot and stick management is more often associated with the arrogance of pure opinion rather than with the considered judgement of a data analyst or a researcher. Data frequently has a sobering effect since it is usually far less clear cut and simple than the would-be decisive manager wishes.

As a teacher, I would prefer, if I did not trust the management, that there was data available to me as well as to the management. This would seem to be a far safer position than to be at the mercy of

opinion unsupported by data. Data empowers the powerless. Of course, the quality of the data must be constantly challenged and we must seek for constant improvement in the quality of the data, but to have no data at all is to leave people vulnerable. Whilst people can lie with statistics, it is, as Dr. Paterson said, the people who lie, not the statistics. I would also suggest that it is very much easier to lie without statistics than to lie with statistics. So even if indicators existed in a punitive system of accountability, I would still argue it was better to have the data than not to have it.

Statistical models. I quite agree that multi-level analysis requires assumptions to be made which have not been subjected to empirical tests. As in much measurement theory, a stable component is identified as 'true', and the unstable as, by implication, everything else including error. Yet the true level of a variable may, indeed, vary. Models built on the assumption that the true component is stable, whilst convenient for labelling purposes, may not conform to reality. It is, therefore, very difficult to choose a single correct model, and there seems to be safety in computing indicators of performance using a number of approaches, and watching all the indicators year by year until we get a sense of which ones conform to the reality of what is actually happening. As Cronbach and Meehle (1955) observed many years ago, construct validity requires a nomological net. Establishing construct validity requires an on-going research effort. The underlying realities of schooling as experienced by staff, teachers, students and parents is what the indicators must be true to, not to some particular statistical model. It is also predictable that indicators will forever be inadequate (cf. Murnane, 1987). It is characteristic of measurement that it contains error, and it is the hallmark of the scientific approach that there is constant effort made to estimate the extent of error in the measurement and to reduce it. Indicator systems need parallel research programmes to work on this kind of issue.

The major issue for Dr. Paterson seems to centre on the reliability of the data. Just how reliable does data have to be before it is looked at by those concerned with the system - - - concerned as teachers or managers?

Dr. Paterson states

> 'the moral case against performance indicators (that they are unfair) finds support in a case from within statistical theory itself. The unfairness arises because, at the level of the individual *school*, statistics can *never* be reliable enough to hold that possibly unique institution accountable.' (emphases added)

This is an incredibly strong assertion. "Never"? So there can be no accountability? There is no such thing as measurably effective or

ineffective teaching or effective or ineffective systems of schooling? Such a view probably betrays a very limited experience of classrooms and is a depressing hypothesis for teachers who strive to be effective. This is not to say that the hypothesis is unsustainable but it certainly is not proven. Furthermore **there is no moral virtue in denying the possibility of accountability** at some level of the system. I would assert, for example, that many inner city schools in the US are miserably failing their students. I would not blame the teachers but *the system* in which there is totally inadequate information. When there is no mechanism for feeding back to teachers what it is possible to achieve, when there is no advocate for students, and no standards are set, the system can spiral downwards towards one which fails the students. If Dr. Paterson had to try to improve such a system I feel sure he would want indicators, even though they might be unreliable in any one year.

The "in any one year" qualification is important and points up a serious inadequacy in the statistical arguments presented. Dr. Paterson never considers the reliability of, say, a three year moving average of indicators. Three repeated measurements may yield highly reliable information even though none of them alone reaches the traditional standard of reliability. In the SOED Standard Tables three years' data is presented. People can see the instability in the data and gradually learn the extent to which they can place confidence in the figures

This brings us to the question of the standards of reliability which should be set...the required level of confidence in the data. Here it must be said that statisticians are frequently guilty of pretending that they can assess 'confidence' without justifying their choice of confidence level. This is what is done when arbitrary levels of statistical significance are espoused...as in Dr. Paterson's paper (e.g. when referring to Strathclyde and Fife data using two standard deviations away from the mean.) The distinction between substantive significance (or magnitude of effect) and statistical significance is vital and is only slowly starting to penetrate into general consciousness in the social sciences. (Glass, McGaw and Smith, 1981; Schmidt, 1992). **In a situation as novel as that of the development of performance indicators in schools we all need to watch the data, relate it to events and withhold judgement until some costs and benefits of Type I and Type II errors can be assessed.**

And here is a nice irony about the democratisation of the data: it could be *teachers* rather than *researchers* who understand most quickly what residuals are important and what residuals represent trivial random variations. Teachers will be watching from the basis of a deep understanding and knowledge of what is going on in their schools and colleges. For researchers to get that same level of understanding would require an enormous amount of collection of qualitative as well

as quantitative data. That would indeed be expensive. I am delighted that the democratisation of data represented by the sharing of indicators between researchers and schools will enable schools themselves to get a feel for the data which will be of great importance in evaluating what levels of significance we should be looking to consider as 'significant'.

The important message in Dr. Paterson's paper (Table 4) is that 100 percent data is to be vastly preferred for indicators rather than data based on sampling. This is what has always been used in the A-level Information System which we run at the University of Newcastle Upon Tyne: we provide Performance Indicators for departments based on as close to 100 percent data as is obtainable (and standard errors next to every indicator).

One hundred percent 'samples' have many advantages, most of all for teachers. As a teacher, I would be irritated by research findings showing gender differences or ethnic differences or social class differences without my being able to ask 'But what about in my school? Are there gender differences there, and if so to what extent are the differences equivalent to those in other schools or are they worse or better? How can I watch them from year to year so that I may know if there is improvement in the way the school is functioning?' If the data is available in fine detail, student by student, people can check the accuracy of that data which pertains to their own practice. Furthermore this student-by-student data addresses the need for an understanding of the variation which will inevitably occur in indicators. When teachers see data student-by-student they become well aware of the way the indicators are built up from the data on each individual student and well aware of the likely fluctuations from year to year - the "errors" or unreliability. Moreover they can attempt to take account of the 'flu, the love crisis, the bereavement and all the many other possible explanations of factors affecting performance.

In short, Dr. Paterson clearly felt that schools could not be given data that was unreliable. The implication was that people in schools cannot learn to look at a mean and consider it as an estimate in which there is some degree of error. I think this position is itself in error, and this was indeed the opinion expressed by head teachers at the CES conference in Edinburgh (August 1992). They clearly expressed their wish to see data and assured researchers that they could interpret it cautiously. Of course, misinterpretation of data is always a problem, whether by researchers or by teachers, or, particularly, by politicians. But at least when there is data available the debate can be joined on the basis of the data, and the error in the data can be readily demonstrated. Any notion of simple minded accountability... such as the most ridiculous: pay tied to indicators...is made more difficult

in the face of the natural variation that one would expect to arise from the many factors that can influence performance.

For a detailed discussion of the extent to which 100 percent data used in an information system can be fair see Tymms (1993) "Accountability: can it be fair?".

Relative ratings. Dr. Paterson presents only two advantages of Relative Ratings (and numbers these as one advantage!): that they take account of pupil ability and are "quite recognisable" to teachers parents and pupils. Here are some other positive features: they can be created from a single set of examination results; they take some account of subject difficulty (and indeed provide clear assessments of this); they are cheap; they deal with outcomes subject by subject rather than with less meaningful aggregates across subjects; I count 6 advantages rather than 1!

Dr. Paterson lists 7 "disadvantages". First the SOED is called 'reprehensible" for introducing Relative Ratings without extensive independent evaluation or the 'testing that occurs through wide use.' Well, they are now in wide use and studies should be made of *how* they are actually used. However, statisticians are perhaps just as reprehensible in having ignored for so long Alison Kelly's excellent work which led to the development of Relative Ratings, work which meets a real need. English and Welsh Examination Boards still wrestle with "subject pairs" analyses when they should probably be using the Relative Ratings approach. Why haven't statisticians shown an interest in this practical problem which has considerable implications for "school effectiveness" studies? If statisticians have not paid attention to the different difficulties between subjects they have missed a major feature of the data. For example, in some English A-level data, if Physics had been made of equivalent difficulty to other subjects there would have been no grades in the lowest category(U), only a 12 percent failure rate ('N' grades) and an A+ category would have been needed. (Fitz-Gibbon, 1991, page 120m). In contrast, there would have been no 'A' grades in English and a failure rate of 28 percent. Differences in difficulty are substantial, not always recognised and need to be considered in studies of "school effectiveness." Why have statisticians tended to miss this feature of the data? Is it largely because they have chosen the school as the unit of aggregation rather than the individual subject? Was this due to the easy availability of school-level data? Is this attributable to laziness or lack of funds? Or to a lack of appreciation of the point Plewis made that children are taught by teachers not by schools?

Another way to look at this unit of aggregation question is in terms of proximal and distal variables. It could be hypothesised that processes

which go on in individual classrooms...variables *proximal* to achievement behaviours... are more likely to have an impact on achievement than management strategies at the school level or any other more *distal* variables such as the type of institution or the LEA. Gray, McPherson and Raffe(1983) in one of the earliest studies of school effectiveness, paid attention to such variables but much work on 'school' effectiveness since then has failed to go to the heart of the educational process. Thus when summarising factors which have been suggested in the school effectiveness literature as suitable for monitoring, Willms' (1992) list does not include teaching and learning processes.

Criticism 2 is that Relative Ratings do not measures progress. This is true, but the fact that height doesn't measure density is not a criticism of height. There have been extensive efforts to make it clear that Relative Ratings do not measure progress. For example, in 'Performance indicators and examination results', I emphasised just this point in a number of ways, including giving graphical representations of value added and relative ratings for two actual schools in Scotland. This Interchange was sent by the SOED to every school in Scotland. The important warning to be noted, about Relative Ratings, concerns a possible negative impact: because in each school the Relative Ratings sum to zero, there are equal numbers of winners and losers - - - a kind of zero sum game is in operation. Relative Ratings therefore tend logically to set departments in competition with each other. This could well be unhealthy for a school. Much will depend upon the management in the school and case studies are now urgently needed of the ways in which Relative Ratings are understood and used. However, it must be acknowledged that Relative Ratings were the only measures available ...for the whole Scottish system ...at Standard grade so there was some excuse for using them. Value added measures between Standard Grade and Highers are to be added to the Standard Tables and are generally preferable when they are available.

Criticism number 3 is: "like adjusted means, (Relative Ratings) refer to aggregates of pupils, and therefore not directly to the individual processes that constitute education." Actually Relative Ratings are not "aggregate data" in the way that "adjusted means" are. In adjusted means (means on means) there is no pupil-to-pupil link in the data. In Relative Ratings there is.

Criticism number 4 is that Relative Ratings "cannot be used to compare schools" This could well be a virtue! One consequence of failing to investigate achievement subject by subject and of allowing the school so often to be the unit of aggregation, is that it has allowed politicians to pose the issue as one of "good schools" or "bad schools" and thus to promote as a solution the setting of schools into

competition with each other. The data do not seem to support the fiction of overall, across-all-departments, good schools and bad schools. Within schools, departments vary considerably and **there is no evidence that the provision of effective education in the country could be improved by closing down some schools and enlarging others.** Monitoring department by department, year by year to ensure a fair chance for all students no matter where they go to school, seems more fair and more promising. This approach can be called equal opportunities or quality assurance - - - whatever you call it, it needs data to be fair.- - - not one-off data, but regular data. The appropriate level of statistical significance is one of the least of our worries and an issue which can only be understood over the years of experience which schools are now beginning to obtain.

Criticism number 5 of relative ratings... that they cannot be used with subjects in which there is not a general ability factor with a strong influence is a very minor criticism and a feature which has been noted before, with a factor analysis showing that Secretarial Studies at Highers did not fit well into the Relative Ratings framework (Fitz-Gibbon, 1991, page 61b) (There is an important point here: should all subjects be assessed in similar ways? Probably not.)

Criticism number 6 concerns the need to take account of gender. The Relative Ratings in the Standard Tables are presented for high and low ability groups and could be also for boys and girls. However, Relative Ratings are certainly not as amenable as regression analysis to including more factors.

Criticism number 7 is another of "it is not all things to all people" type of criticism. True, the method is not applicable to truancy but the problems with assessing truancy are not statistical. As a Head said to me "It's quite simple: now we have to report truancy rates for the press there won't be any more truancy; only excused absences."

To conclude this section on Relative Ratings: they are an area in which much further research could be undertaken ... the most important being how they are used and the impact they are having if any ...but, nevertheless, the production of this information represents a quite remarkably open and detailed system which has managed to provide data at the departmental level for almost every school in Scotland.

Student level regression. Dr. Paterson lists as a disadvantage of student level regression 'It is expensive'. Data have to be collected on students that would not be routinely available'. What constitutes expensive will always be open to debate, but I think it can hardly be considered expensive to produce value added scores between standard grade and Highers. It represents about three weeks' work for one researcher in the Scottish Exam Board because they have all the data to hand and it is simply some extra processing. The major problem is the matching up Higher grades to Standard grades, student by student, but even that problem could be rapidly overcome

by use of a student number system, which could also improve the extent to which the papers are marked 'blind', a practice which would be more proper than the current practice of leaving students' names and centre names on the paper. Students' names and centre names will convey gender, ethnicity, social class, religion and may, therefore, introduce bias. The Scottish Exam Board along with all the English Boards, should improve practice in this area. Dr. Paterson might respond that the calculation of value added from standard grade to Highers represents only one year of work and that the generation of value added data (based on student level regression) for standard grade would require a test given earlier. This is true and to the extent that such tests cost something the system will have that expense associated with it.

Dr. Paterson might argue that the use of a prior measure of achievement or aptitude is not sufficient, that one must take into account social circumstances. But social circumstances are expensive to measure accurately and, indeed, Dr. Paterson himself is one of the experts in this area (see Paterson, 1992). However, it is not at all clear that an indicator system should offer the excuse of home circumstances. The effect of social class and home background should continue to be monitored, but what schools need to know is, given children with certain *developed* abilities are the departments producing an appropriate set of exam results for those students? As a parent in an inner city deprived area, I would want to say if I send you a child with certain developed abilities, I expect that child to come out with the same exam results as a child with the same level of developed abilities from a more privileged home. To build social class into an indicator system is to excuse social class differences in the progress made in school. We must watch for such differences and if they are located, we need an indicator system to find if there are schools overcoming those differences and, if so, what they are doing. We need an indicator system to monitor whether, if we put more resources into schools in under-privileged areas, we can improve the progress made by students, and how best those resources are spent.

In other words, the use of performance indicators *can* guide public policy and help us to locate effective practice. Such data is worth some expense but, equally, a very simple system of feeding back data based simply on a prior measure of achievement linked to externally-set, blind-marked examination papers provides excellent feedback for teachers on a student by student basis indicating the achievement of each student in the examination subject, taking account of the difficulty of the subject that year with that exam board, and the students' prior level of attainment. Such data is viewed by teachers with great interest and can hardly be said to be expensive. Indeed, many schools already have prior measures of achievement which they wish to use for their own purposes in the identification of learning disabled or very able children. The systematic collection and use of such data among very large consortia of schools is all that is needed

for the production of quite good performance indicators, indicators which can have multiple uses.

Dr. Paterson's second objection to student-level regression analysis is that the method is difficult to understand: 'the problem is whether people are happy to accept a method in which the implementation of the principles (of regression analysis) is not so transparent as is the method of adjusted means'. I can only say that a simple line of best fit is well understood in my experience and, indeed, there is not a single school in which a maths or science teacher would not be available to explain the process to anybody wishing clarification. Multi-level modelling is a different proposition.

Dr. Paterson's third point relating to regression analysis is that the regression line itself may have been influenced by a general undesirable pattern in the data such as low expectations as regards performance of girls. There is no denying that a regression analysis may fail to lead to the detection of a systematic effect throughout the country but that highlights one of the values of an information system which provides department-by-department data. If no department is removing the gender difference, should any be held accountable for it? It is this sorting out of what Deming (see Neave, 1990, p.264) calls 'common cause' variation' from 'specific cause', the sorting out of what is general in the system from what is unique to individual schools which is a major contribution that performance monitoring can strive to provide.

In passing, it is worth noting that the notion of teacher expectations is a much overworked notion, a notion particularly attractive to administrators who may well be aware of the tones in which the original research was reported 'Nothing was done directly for the disadvantaged child at Oak School. There was no crash program to improve his reading ability, no special lesson plan, no extra time for tutoring, no trips to museums or art galleries....."' (Rosenthal and Jacobsen, 1968 pp 182). In other words, it was cheap and only involved blaming the teachers. Anyone wishing to be disabused of this notion that teacher expectations have a strong effect on student achievement should read Elashoff and Snow (1971), a meta analysis of the effects of teacher expectations (Raudenbush, 1984) and Teddlie, Stringfield, Wimpelberg, & Kirby, (1989) on Effective Low SES schools in Louisiana. The original work on teacher expectations was something of a Watergate in research: The intake measure would have classified a third of the entering class of the middle-class primary school as retarded. The modal score on the reasoning component of the test was zero IQ. It is sad that this blame-the-teacher, putative research finding is so widely popular.

Dr. Paterson states that a 'very fundamental problem' stems from statistical uncertainty which is unavoidable even with 100% samples

from each school. However, Dr. Paterson's arguments in this section seem to stem from an attempt to classify schools 'Only approximately the top and bottom 5% of schools can be reliably distinguished from average in any particular year in the sense of their school effects being more than two standard deviations away from the mean school effect'. Classifying 96 out of 100 correctly seems pretty good though the point of classifying escapes me. It is important to notice that in the ALIS project we do not speak of 'school effects'. From the beginning the decision made was to use the statistical term 'residuals' so as not to over-state the interpretation.. Schools understand that *residuals* are *what is left over* after the student intake has been taken into account. This is not an impossibly difficult concept, not in the least. By the use of the technical term 'residuals', we emphasise that the residual is simply what is left over, and that will include, possibly, the effect of the teaching in the department, but also, possibly, such factors as text book, exam board, influence on the atmosphere in the class of one particular student, the incidence of glandular fever that year in that school, the time at which the course is taught (six periods of maths on a Friday may not be the best distribution of time for learning mathematics). There are many other factors that may enter into the residual, plus a good component of error.

One further point is that schools do not look at a single set of data nor is classification necessarily a use to which data is to be put. There is undue emphasis given to this straw-man problem. With experience we might be able to judge whether three years is a sufficient buffer against unreliable judgements in particular indicator systems. Indeed, perhaps only three-year moving averages should be reported to, for example, inspectors or governors.

Accountability in national policy Dr. Paterson expresses dismay at the lack of accountability of policy makers and politicians. Actually, if indicator systems were in operation policy makers would be much more accountable than at present because they would find it more difficult to make unsubstantiated claims and the results of their latest initiatives would quickly come on stream. The option of leaving data unpublished, such as seems to have been taken with respect to some TVEI data (as noted in Fitz-Gibbon, 1990) , would not be available. Indeed, the Relative Ratings data provides two examples of such possibilities of accountability.

First, the fact that Relative Ratings have been found to correlate quite highly (about 0.70) with value added scores (See Fitz-Gibbon, 1991 p76, and Dr. Paterson's Table 3) **would not be possible if there were overall "good schools" and overall uniformly "poor schools'.** Were a school to have uniformly positive residuals (i.e. uniform measures of relative progress across all subjects) the correlation with relative ratings, in which the indicators must sum to zero, would be theoretically zero. The fact that the correlation is strong is an

important finding from the Scottish system and supports our experience in the ALIS project: It is departments more than in schools, in which there are important variations in outcomes. **This finding undermines the notion of using indicators to promote competition between schools and supports the preferable approach of maintaining quality in every department in every school**. Indeed, this department by department data is what schools are interested in. League tables are simply a distraction from the need to ensure quality in every institution, which is the important task. As the industrially-oriented jargon of 'value added' or 'quality assurance' is introduced into education it is important to note that these terms mean essentially the same as the more familiar "equal opportunities".

Secondly an interesting interpretation can be postulated for Dr. Paterson's demonstration in Table 3 that the correlation's between Relative Ratings and pupil level regression indicators have fallen from 1984 to 1991. This drop implies that in some ways schools have become more homogeneous. Perhaps schools are becoming more differentiated: some schools becoming more uniformly effective and some schools becoming more uniformly ineffective... or it could, possibly be that all schools are becoming uniformly effective with little variation. The first possibility seems more likely and could be an outcome of increasing parental choice, and/or of increasing disparities in funding. Is the operation of the market tending to produce niches? Is there any corroborative evidence from other sources? Is it in anyone's interest to have a low end of the market in education? Is it not just storing up problems for the future? Who, then, is responsible for policies which might be having that effect? We can't be sure without further data, but we need to consider the change in the correlation with Relative Ratings as possible evidence ... as a possible indicator of national changes.. A change of government (and) policy would be very helpful in casting further light on the issue: can the increasing discrepancy be turned around by other policies? In this kind of way monitoring systems *can* help to make policy makers responsible. For another example of indicators informing the evaluation of policy, and, therefore, the accountability of politicians, see Tymms' study of the Assisted Places Scheme (Tymms, 1992). It provides a fine example of the way an indicator system can be used to illuminate national policies.

How should indicator systems be run?

Answer: emf: empower, monitor and feedback. Data on school effectiveness should not be reserved to researchers in the hope that they will come up with some eternal verities from which schools can learn. I would not advise holding one's breath or delaying decisions pending the emergence of a "grounded theory of professionalism." Education is inevitably a fairly complex and, possibly, chaotic process. The way to deal with such processes is to monitor carefully

and to watch for wild fluctuations, and see if one can set them right. Gene Glass hinted at this idea some years ago before chaos became so popular (Gleick, 1988; Ruthen, 1993) when he suggested that what education needed was not research but a fire fighting methodology of watching for problems (Glass, 1979). Schools should be empowered with a good deal of decision making, monitored by agreed methods on agreed outcomes and should receive feedback on the basis of this monitoring. Empower, monitor and feedback would seem to be a good formula for the management of a complex system (Fitz-Gibbon, 1992).

I believe that much of Dr. Paterson's opposition to performance indicators arises from his fear that they will be used in a domineering culture of management, and much of my support for performance indicators stems from my belief that data empowers people, promotes a spirit of rational investigation, and provides an important bedrock for systems which are fair to staff, students, subjects and society. It is my belief that education is more in danger from the arrogance of unsubstantiated opinion than from management moderated by good quality data.

REFERENCES

Aitkin, M. and Longford, N. [1986] Statistical modelling issues in school effectiveness studies. Journal of the Royal Statistical Society. Series A.149(1) pp 1-43

Cronbach, L.J. and Meehl, P.E. (1955) Construct validity in psychological tests. Psychological Bulletin. 52, May.

Elashoff, J.D. and Snow, R.E. [1971] Pygmalion reconsidered. New York:Wadsworth Publishing Co.

Fitz-Gibbon, C.T. (1990) Learning from unwelcome data: Lessons from the TVEI examination results. In Hopkins, D.(Ed.) TVEI at the Change of Life Clevedon,Avon: Multi-lingual Matters.

Fitz-Gibbon, C.T. (1992) The design of indicator systems, the role of education in universities, and the role of inspectors/advisers: a discussion and a case study. Research Papers in Education. 7(3) pp271-300

Fitz-Gibbon, C.T. [1991] Evaluation of School Performance in Public Examinations A report for the SOED. University of Newcastle Upon Tyne: CEM Centre

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

Glass, G.V. McGaw, B. and Smith, M.L. (1981) Meta-analysis in social research. Beverly Hills: Sage Publications.

Gleick, James (1988) Chaos: making a new science. London: Heinemann, 1988

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.

Murnane, R.J. (1987) Improving education indicators and economic indicators: the same problems? Educational Evaluation and Policy Analysis 9(2)pp 101-116

Neave, H.R. (1990) The Deming Dimension. Knoxville; Tennessee: SPC Press, Inc.

Paterson, L. [1992] Socio-economic status and educational attainment: a multi-dimensional and multi-level study. Evaluation and Research In Education

Raudenbush, S. (1984). Magnitude of Teacher Expectancy Effects of Pupil IQ as a Function of Credibility of Expectation Induction. A Synthesis of Findings from 19 experiments. Journal of Educational Psychology 76(1), (pp. 85-97).

Rosenthal, R. and Jacobsen, L. [1968] Pygmalion in the Classroom: Teacher Expectations and Pupils" Intellectual Development. New York: Holt, Rinehart and Winston Inc

Ruthen, R. (1993) Adapting to complexity. Scientific American. Jan. pp110-117

Schmidt, F. L. (1992) What do data really mean? American Psychologist. 47(10) pp 1173-1181

Teddlie, C., Stringfield, S., Wimpelberg, R., & Kirby, P. (1989). School Effectiveness and School Improvement. In B. Creemers, T. Peters, & D. Reynolds (Ed.), Second International Congress for School Improvement and School Effectiveness, Rotterdam: Swets & Zeitlinger.

Tymms, P.B. (1992) The Relative Success of Post 16 Institutions in England (Including 'Assisted Places Schools') British Educational Research Journal 18(2) pp175-192

Tymms, P.B. (1993 In press) Accountability: can it be fair? Oxford Review of Education.

Willms, D. (1992) Monitoring School Performance. A guide for Educators. Washington D.C. London: The Falmer Press