Chapter 11

# Performance Indicators, Value Added and Quality Assurance

Carol Fitz-Gibbon

One of the more important tasks for managers in the 1990s will be to collect systematic information about the performance of the units that they are managing. In other words management in the 1990s almost certainly implies the setting up of performance indicator systems. These systems will take time and cost money, because the collection, analysis and interpretation of data are exceedingly time consuming activities. This will be time and money well spent only if the performance of the system is improved by the existence of the performance indicators. It follows that one of the most important characteristics of performance indicators should be that their impact on the system should be beneficial. Whether or not the impact is beneficial may rest crucially on two actions taken by management: the particular indicators chosen and the manner in which they are used.

## MEASUREMENT GIVES MESSAGES: THE IMPORTANCE OF CHOOSING THE RIGHT INDICATORS

The choice of performance indicators represents a signal from management as to the features of the system that are of most concern. Consider the tragedy of the cross-channel ferry called the *Herald of Free Enterprise*. It might well have been the case that the time taken to cross the channel and return was carefully monitored, because of the need to adhere to a strict timetable and to make as many crossings as possible, so that each crossing represented a maximum intake of money into the system. This monitoring would have acted as a signal to operators to concentrate on a rapid turn-round in each port. If, concurrently, there was no systematic monitoring of what might be called 'near misses', such as leaving port with the bow doors open, then less attention would be paid to that feature of the system than to the speed of turn-round. The consequences of such monitoring could be disastrous in terms of safety. The design of beneficial performance indicators is surely one of the most onerous and demanding responsibilities placed on management.

## THE MANNER OF USE OF INDICATORS

Not only must the choice of indicator be made carefully but the conditions of reporting of the indicator may also be vital in determining the effect it has on the system. In the airline industry, for example, pilots can report 'near misses'; that is, situations in which, although no accident happened, planes were flying too close for comfort or some evasive action was needed to avoid an accident. Pilots do not need to give their names, so that 'near misses' are not held against them. If the system required that pilots gave their names then it could be seen in terms of punitive surveillance and the self-report aspect would be in jeopardy. Since the report of 'near misses' can only be made by those aware of these incidents, it is exceedingly sensible to make sure that such information *is* collected, rather than to set up a system in which data would simply not be produced or would be corrupted.

The philosophy behind the confidential reporting of 'near misses' is that these potential accident situations arise from *faults in the system rather than faults attributable to individuals*. Is this a principle that can be applied generally to complicated systems, such as airlines, education or cross-channel ferries? W. Edwards Deming believes it is, suggesting that most inefficiency is due to defects in the system rather than indolence or unwillingness on the part of employees. If the system is at fault, the system needs monitoring in order to set it right. The monitoring must be based on good data and that often requires some confidentiality so that the temptation to make the data simply *look* good is not built into the system. Highly punitive surveillance systems simply become corrupted. This was evident in Eastern Europe in a widespread manner, with phoney targets and phoney feedback of the extent to which targets were being met. The consequence was an economic system full of disinformation and exceedingly ineffective.

## EXAMPLES OF INDICATORS IN USE IN EDUCATION

We need, then, to set up systems of performance indicators in which the indicators are chosen so that their use produces a beneficial impact on the system. This requires a careful choice of indicators and of the methods by which they are reported and used. With these considerations in mind, let us consider some performance indicators that are recommended for use in education in such documents as the Coopers & Lybrand report, *Local Management of Schools* (Coopers & Lybrand, 1988). One kind of indicator is the percentage of students achieving a pass or a particular grade. In its worst form this indicator may be the 'percentage pass rate', measured by the percentage of those entered for the exam who obtain a passing grade. The impact of a percentage pass rate indicator is unfortunate in a number of ways. Logically, percentage pass rates should be expected to encourage teachers to push out of their classes students who are likely to fail, whether or not it is in the student's interest to be encouraged to leave the class or not. Logically, a percentage pass rate should encourage teachers to concentrate their teaching on the borderline students. The students heading towards an A or a B are unlikely to end up failing and can therefore be neglected. The teacher should concentrate on improving the pass rate by ensuring the

borderline candidates get through the examination. If these logical approaches to 'fixing' the indicator were adopted the result would be unfair practice: pupils would not be treated equally. The consequence for some students of failing to obtain their As may be as grave as for other students of failing to pass the examination at all. To encourage fair practice, performance indicators must be chosen so that each pupil counts equally in the computation. If the percentage reported as an indicator is not of those entered for the examination but of the year group there are still unfortunate implications in the publication of percentages attaining various grades. In colleges, where enrolment is fluid, they face the problem of wanting enrolment for the money but not wanting students who are going to fail and damage the performance indicators.

Of course these problems can be exaggerated. If the drop-out rate is also monitored, the temptation to push students out of a course as the exam draws nearer will be moderated by a desire not to worsen the drop-out indicator. These considerations suggest that to monitor a system adequately a fair number of performance indicators will be needed. Indeed, the use of a large number of indicators may be fairer and more beneficial in the long term than concentrating on one or two indicators. There is always the danger, however, that only a simple indicator like the percentage pass rate will be reported in meetings or in the press. It would therefore be wise to choose a good single indicator which can be emphasized and the use of which would be least damaging. The kind of value-added indicators used in the A Level Information System (ALIS) project, and widely used in research on school effectiveness, treat each pupil as equally important. By also taking into account the most important influences on the examinations, the indicators are probably the best single indicators available.

Another example of an indicator that could have unfortunate effects is the relative rating used in Scotland (Kelly, 1976; Fitz-Gibbon, 1992). The relative ratings provide comparisons between the performance of different departments within the same institution. The relative rating indicator takes account of the two dominant influences on examination results in each subject: the difficulty of the examination in that subject and the general aptitude of students enrolled in the course. This system is a sophisticated development of a practice used in many schools, that of looking at the grades in one subject in comparison with the grades the same students attained in other subjects. While it is an interesting and sophisticated statistic the impact could be unfortunate in that it places departments in competition with each other. The maths department will be placed in competition with the physics department because they share the same students. The maths department should logically load students with maths homework to encourage them to neglect their physics, because if physics performance declined the maths department could get a better relative rating.

An interesting feature of relative ratings, however, is that they tend to correlate about 0.70 with measures of 'value added'. Value added indicators show the extent to which students are gaining examination results relatively better or relatively worse than similar students elsewhere studying the same subject. In other words, departments which are *relatively* effective in their school tend to be those which are effective *in comparison with other departments in the same subject in other schools*. Value added indicators compare

maths departments with other maths departments, physics departments with other physics departments and so on, in all the institutions participating in the performance indicator system. It should be noted that if schools were uniformly good or bad the correlation of value added with relative ratings would be close to zero. The high correlation illustrates a point made many times in the ALIS project: departments differ substantially even within the same school.

Now that the Scottish Examination Board, in conjunction with the Scottish Office Education Department, is producing relative rating indicators for just about every department of every school in Scotland, it will be very interesting indeed to find out how this information is used in schools. Here is a prime need for careful qualitative research. How do headteachers interpret the data? To whom do they communicate the data? With what messages are the data communicated? Do those to whom the data are communicated interpret the messages correctly, misinterpret the data, accept the data, find them interesting, find them useful, find them threatening? The same questions need to be asked about the reception and use of the ALIS data.

## DESIGNING MONITORING SYSTEMS

We have considered up to now existing indicator systems, but in many situations indicator systems do not as yet exist and must be created. The creation of indicator systems must obviously take account of the need for the indicators to have a beneficial impact.

In the first instance, decisions have to be made about which outcomes need to be monitored. A system in which the outcomes are not monitored is flying blind, totally unable to set reasonable targets or know whether it is doing well or doing badly. *The measurement of outcomes is absolutely fundamental to the monitoring of the performance of a system.* The outcomes that are measured must be chosen carefully. Clearly they must be outcomes over which the system has some influence. This may seem obvious but it has been suggested that delinquency rates should be collected school by school. This would be reasonable if it had been shown that schools had an influence on delinquency rates. It *has* been shown that delinquency rates vary from school to school, but so does the incidence of leukaemia. Until the causes or mechanisms are understood it would seem unfair to regard staff as responsible for delinquency rates in their school. 'No accountability without causality' might be a rallying cry.

An example of an indicator that may be susceptible to some distortion is the truancy rate. As one head remarked, commenting on the requirement to publish truancy rates: 'It's quite simple: from now on we won't have any truants. It's all a matter of definition.' Another corruptible indicator may well be the results of Key Stage assessments, assessment done by teachers and reported so that, in this era of competition, we can see how well they are teaching. The extent of checking required to verify testing conditions and marking standards make the Standard Assessment Tasks (SATs), as proposed, unworkable except as a system resting on 'scout's honour'. In a framework of competition and unemployment, grave misgivings about the validity of publicly reported SAT results would seem to be reasonable. Confidential use of well-tested items from the

Assessment of Performance Unit might have provided valued feedback to teachers, but the proposed system of public reporting of teacher-marked performance lacks credibility. It is a system that also provides little in the way of safeguards against bias. (This discussion of SATs was written about a year before their collapse and their subsequent revision by Sir Ron Dearing.)

In the measurement of outcomes many managers and researchers look at whole-school indicators, such as the sum total of examination results produced by a year group. This aggregate sum may be of passing interest but the managers will need to know the contributions made by each department. Furthermore, it is the departments that are the units that are managed to provide the education that results in the examination passes. The departments, therefore, need indicators.

There is much reliance currently placed on the provision of raw data. League tables are drawn from information on the number of students achieving various grades, which must now be published. We need to repeat, again and again, the following idea. That is, we need to teach this idea to politicians and to the public.

## FOUR KINDS OF DATA

There are four kinds of data.

*Raw data* are simple to understand, but often almost impossible to interpret. For example, if we know that 10 per cent of students attained an A at GCSE level in English we know what this information means but we do not know what it implies. The attempt to evaluate or interpret this information often leads to a second kind of data.

*Comparative data.* The 10 per cent As may be compared with the national rate at which As are attained. This adds to the body of information, but interpretation and evaluation of the data are still very difficult. The comparison with a national average raises the question: Is it reasonable to expect this school to attain results in line with the national average? Or should its results be better? Should its results be worse? This kind of question leads on to the third kind of data.

*Residuals* or adjusted data. This kind of data could also be called *fair comparative* data or *contextualized* data. Although it represents an intrusion of what might be seen as statistical jargon, I will use the term 'residuals' for this third kind of data because the term itself carries important implications. A residual is defined as the difference between the result obtained and the result predicted from measurements of factors known to be correlated with the outcomes. The results that could reasonably have been expected in an examination, for example, can be predicted from a knowledge of the pattern of results across many institutions in that subject, for that year, and a knowledge of the intake characteristics of students. The residual is the difference between an actual and a predicted grade. The computation of 'residuals' basically is the calculation of what is left over after taking account of important factors which influence the results and over which the school and staff have little control. The school and staff have little control over the prior achievement, ability and interest of

students. If these are measured two years before an examination, say, they form an important baseline against which to judge the examination results. The change from two years earlier is often called the value added and the techniques of measuring value added are the techniques of obtaining residuals (a better term than value added would be relative value added). The argument is then made that the performance indicator which best represents the examination effectiveness of a department is the average residual obtained in the department. If the average residual is positive it implies that students on the whole obtained higher grades than would have been predicted on the basis of their prior achievement and other factors that have been taken into account. In the A Level Information System (ALIS), for example, we also take into account home background and scores on a specially administered aptitude test (the International Test of Developed Abilities). Gender is also an important factor in performance and seems to have systematic effects across the A level results.

Residuals are as close as we generally get to fair performance indicators but it must be emphasized that the residual is *what is left over* after certain factors have been taken into account (to the extent that they can be taken into account) by the measurements available in the system. There are many factors influencing the outcomes that are not measured and are not, therefore, taken into account. The effect of some of these factors will be present in the residuals, and the effects of the errors of measurement will also be present in the residuals. Only a part of the residual can be seen as indicating the 'effect' of the department on students' performance in the examinations. A fundamental principle in measurement is that all measurements contain error and some estimate of this error is what makes a measurement scientific. In the case of residuals the amount of error involved in their estimation is computable and indeed in the ALIS project we report average residuals for departments with an error indicated alongside as a warning against the over-interpretation of small differences.

*Experimental data.* There is a fourth kind of data, which is rarely available in the system. If it were available it would provide the most conclusive and the most fair evidence of effectiveness. The fourth kind of data arises not from surveys, passive observation of the way the system is working with all the problems inherent in the built-in self-selection mechanisms in various courses and various schools, but from randomized experimental assignment. In other words, if we ran clinical trials, as is done in medicine, the resulting data would be the Gold Standard Data, those derived from controlled experiments. It is sometimes said that experiments are impossible in education. This position is far from accurate. There have been controlled experiments, yielding very important cost-benefit analyses of various interventions, most notably in the area of early childhood education (Lazar *et al.*, 1977). Furthermore, the ideal of undertaking 'reforms as experiments' (Campbell, 1969), in order quickly and accurately to evaluate their impact, must eventually be adopted if progress in social science is to be sufficiently rapid and accurate to achieve the kind of society that we would like for our grandchildren. However, this issue of experimentation cannot be further explored in the confines of this chapter. It is simply an issue that must be raised again and again until people, especially policy-makers, begin to recognize and understand the implications. The classical best-seller in

educational research, Campbell and Stanley's *Experimental and Quasi-experimental Designs for Research* (1966), should be a required textbook in all research methods courses.

How, then, should educational managers approach the design of performance indicator systems? The system needs to be designed in the framework of a rationale. Goals have to be identified, and factors that have an impact on those goals also need to be identified so that we have the basis for producing fair performance indicators or residuals. Achieving these steps represents much, but there is one further important step to take: to add process variables in order to investigate, through the system, possible ways of achieving improvements.

To illustrate, rather sketchily, the procedures adopted in developing an indicator system, I will outline three systems: the A-Level Information System, a monitoring system for BTEC and a year 11 indicator system. For each one we need to consider the outcomes, the covariates that predict those outcomes and that we need to take into account in order to make the comparisons between outcomes fair, and the processes that are of interest in their own right or in order to examine their effects on the outcomes.

## THE A LEVEL INFORMATION SYSTEM (ALIS)

### Outcomes/goals

Consensus on goals is not difficult to achieve among A-level teachers and managers. Five major outcomes are monitored yearly in the ALIS project: examination results, students' attitude to the subject, students' attitude to the institution, students' aspirations *vis-à-vis* higher education and participation in extramural activities (the last being an indicator of the quality of life in the sixth form). These goals reflect the notion that in addition to getting reasonable examination results teachers hope that students enjoy their time in the sixth form and have a broadly educational experience.

### Covariates

The factors which are taken into account in evaluating these outcomes are prior achievement, gender, ethnicity and socio-economic status. It is found in general, however, that as a matter of empirical fact the socio-economic status of students adds little to the prediction of A-level grades when there is some measure of prior achievement available. In any case, adjustment of indicators for socio-economic status is also problematic in that it implies that less is expected of equally able students if they are from poorer backgrounds. This is not a desirable message. In the ALIS project, therefore, the major indicator is based on the prediction of A levels from GCSE results. Additional tables taking account of other factors are available but the main table rests on this simple measure of value added.

### Process variables

In addition to these outcome variables and their covariates the data collected from students by questionnaires contain students' estimates of the frequency of

use of various teaching and learning activities, such as dictating notes, working in pairs and presenting work to the class. In order to generate ideas about the kind of teaching that might be effective these process variables are related both to students' outcomes on examinations and to students' attitudes to the subjects.

## THE BTEC INFORMATION SYSTEM

### Outcomes

The BTEC-awarded grades and measures of satisfaction similar to those in the ALIS project are the outcomes monitored. It might be thought that qualifications provided by the Business and Technician Education Council would not relate to prior achievement in academic subjects such as GCSE. It is true that the correlations between GCSE and BTEC grades are weaker than those between GCSE and A levels but they are far from negligible and too large to be neglected when looking at the performance of *groups*, as opposed to *individuals*. The correlations between GCSE and vocational qualifications reflect the general finding that prior achievement measures representing general aptitude are good predictors of performance in a variety of situations, including in jobs (Hunter and Hunter, 1984) and in obtaining vocational qualifications.

### Covariates

These are prior achievement, gender, socio-economic status and ethnicity. A major feature in the BTEC Information System lies in the *process* variables assessed on the questionnaire. BTEC has developed careful quality assurance procedures in terms of the provision of resources, time, materials and computers to be made available in institutions running their courses. The questionnaires obtain students' perceptions on the adequacy of the resources available on particular courses. This provides interesting and important comparative feedback for course providers and for BTEC moderators.

## A YEAR 11 INDICATOR SYSTEM (YELLIS)

The outcomes of concern in year 11 are somewhat different from those at A level. Examination results (GCSEs) are of course a matter of prime concern. In addition, schools wish to know if students feel safe at school and, if not, where it is that they do not feel safe. Schools wish to know if students are responding well to target setting and individual action planning, and other such procedures urged on teachers. As at A levels, schools are interested in students' responses to various subjects and the extent to which they enjoy the school experience, get along with staff and get along with other pupils. Other important outcomes relate to career choices and aspirations. Covariates present a problem as there are no examinations prior to GCSE. We have used Raven's Standard Progressive Matrices and, also, a maths test and vocabulary test. These are used along with measures of 'cultural capital' (Bordieu and Passeron, 1977), gender and ethnicity. By monitoring these outcomes schools may be able to improve. At least the monitoring will make them aware if the situation starts to deteriorate. It may also be found that some schools are producing better outcomes than

others, in which case studies ought to be made of the management practices, resource allocations, structural arrangements, ecology of the building and other aspects that may contribute to the positive outcomes.

## WE'VE ONLY JUST BEGUN

Performance monitoring will be the growth area in the 1990s. This was predictable from the steadily increasing availability of computers and, indeed, education as a discipline comes late into an era of performance indicators. Such indicators have been used for some years now in other public services (or what were previously public services), such as the water industry, the railways and hospitals. We need, as a profession, as researchers, as managers, as policy-makers, to work together to produce indicator systems that improve education rather than cause damage. We need to monitor the monitoring.

## REFERENCES

Bordieu, P. and Passeron, J. C. (1977) *Reproduction: In Education, Society and Culture*. London: Sage SES.

Campbell, D. T. (1969) 'Reforms as experiments', *American Psychologist*, 24, 409-29.

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

Coopers & Lybrand (1988) *Local Management of Schools*. London: HMSO.

Fitz-Gibbon, C. T. (1992) *Performance Indicators and Examination Analyses* (Interchange number 11). Edinburgh: Scottish Office Education Department.

Hunter, J. E. and Hunter, R. F. (1984) 'Validity and utility of alternative predictors of job performance', *Psychological Bulletin*, 96(1), 72-98.

Kelly, A. (1976) 'A study of the comparability of external examinations in different subjects', *Research in Education*, 6, 50-63.

Lazar, I., Hubbell, V. R., Murray, H., Rosche, M and Royce, J. (1977) *The Persistence of Pre-school Effects: A Long Term Follow-up of Fourteen Infant and Pre-school Experiments*. Ithaca, NY: Community Service Laboratory, New York State College of Human Ecology.