CMPO Working Paper Series No. 03/073

The Use and Usefulness of Performance Measures in the Public Sector

Carol Propper[1]
and
Deborah Wilson[2]

[1]CASE, LSE, CEPR and CMPO, The University of Bristol
[2] Department of Economics and International Development, University of Bath and CMPO, The University of Bristol

May 2003

Abstract

The paper focuses on the empirical evidence on the use and usefulness of performance measures in the public sector. It begins with consideration of the features of the public sector which make the use of performance measures complex: the issues of multiple principals and multiple tasks. It discusses the form that performance measures may take, the use made of these measures and the responses that individual may make to them. Empirical examples from the fields of education and health, with a focus on the US and UK, are examined. There is clear evidence of responses to such measures. Some of these responses improve efficiency, but others do not and fall into the category of 'gaming'. Generally, there has been little assessment of whether performance measures bring about improvements in service. The paper ends with consideration of how such measures should be used and what measures are useful to collect.

Keywords: benchmarking, public sector, performance measures

JEL Classification: H4, J3

Acknowledgements

We thank the Leverhulme Trust for funding this research and an anonymous referee for very helpful comments on a previous draft..

Address for Correspondence
Department of Economics
University of Bristol
12 Priory Road
Bristol
BS8 1TN
Carol.Propper@bristol.ac.uk
D.Wilson@bristol.ac.uk

CMPO is funded by the Leverhulme Trust.

1. Introduction

Performance management in government has received increasing interest since the late 1980s, fostered by the 're-inventing government' movement (Osbourne and Gaebler 1992). The American Government and Performance Results Act of 1993 required all federal departments and agencies to develop 5 year strategic plans linked to measurable outcomes, via a series of annual performance plans from 1999 onwards. These performance plans had to cover each programme activity set forth in the agency budget, with specific performance measures, and objective, quantifiable and measurable goals (Kravchuk and Schack 1996). The Job Training and Partnership Act of 1982 (JTPA) was the first large scale, federally funded programme to mandate the use of performance measures in state and local programmes. They have since been introduced in welfare to work programmes, child welfare agencies, child support enforcement programmes, and other programmes partly or wholly funded by the government. Under many of these programmes agencies are required not only to publish performance measures but to link payments to them. The Job Training Partnership Act is one well studied example, another is the Temporary Assistance for Needy Families Programme, under which states have competed for large bonuses based on their success in meeting employment and teenage birth reduction goals (Heinrich 1999).

In the UK, the Financial Management Initiative, introduced in the early 1980s, embodied performance management but was assessed as being unsuccessful in influencing the allocation of public sector resources or increasing the degree of public accountability (Osbourne et al 1995; Sharifi and Bovaird 1995). While some performance measures were part of the Conservative adminstration's management of the public sector, there has been a large rise in their use following the election of the Labour administration in 1997. Performance targets, their publication and the linking of such targets to the resources allocated by Treasury to government departments is now widespread in the UK public sector. In addition, individual providers of public services now may get specific rewards linked to their performance with respect to such targets. Some of these rewards are non-monetary, for example hospitals that reach certain thresholds of performance are to get 'earned autonomy'. Some indirectly reward good performance, for example schools that perform well get more pupils. Finally, some are direct financial rewards for performance, for example the team based bonuses being piloted in Customs and Excise and Job Centre Plus (Makinson 2000).

Much has been written in theory about performance measures. However, less is known in practice about their operation, particularly in the public sector. In addition, even fewer studies have examined whether performance measures actually achieve the aims of the programme or government agency, although this is ideally the question that should be asked: do performance measures help agencies achieve the goals they have been set by policy makers? Given the lack of evidence on this topic, the focus of this review is somewhat broader. We begin with a discussion of the key features of public services that affect the use and usefulness of performance measures in the public sector. We then provide a description of what performance measures are and the different ways they may be used. We review the evidence on the use of performance measures in three key public service areas: education and job training, health and education. The evidence in the first area comes almost entirely from the

USA. We examine this evidence as the large number of studies of performance management in this area means that researchers have come closest to asking and answering the question of whether performance management improves performance. The other two areas have been selected because performance management is important and growing in both sectors in the UK. For these two sectors we review both UK and American evidence. This is both because these countries are seen as leaders in the use of performance management in the public sector, and because, as a consequence, the literature in this field has a predominantly UK and US focus.

2. The special features of the public sector

Performance management is used in both the private and public sector. While many of the issues that arise in its use are common to both sectors, researchers studying the behaviour of public sector organisations have recently drawn attention to the fact that the public sector is different from the private sector and therefore a public sector organisation faced with a change in incentives will not necessarily behave in the same way as a private sector one [1].

Dixit (2002) stresses two important features of the public sector. The first is that bureaucrats often serve several masters: these may include the users of the service, payers for the service, politicians at different levels of government, professional organisations. The second, a consequence of the first, is that the agency and so the bureaucrats who work in it often have several ends to achieve. For example, they are often expected to increase both efficiency and equity in the delivery of public services. These features are termed multiple principals and multiple tasks (or goals) respectively. These two characteristics mean that the incentives provided in the public sector should be less high powered than for the private sector.

Even if we consider just one principal, there may be divergence from the predictions of standard agency theory in the public sector context. First, as Dixit points out, the standard agency problem assumes a risk neutral principal. However, at least one group of public sector principals – elected politicians – cannot 'diversify' the effects of bad outcomes. This may make them very risk averse, which again means that the results from the standard theory will not apply. A second diversion from standard theory concerns the issue of moral hazard on the principal's side. For example, an individual may decide to make her career within the public sector, and make career-specific investments to do so, only to find that her pecuniary and/or non-pecuniary benefits are worse than expected. Hence, ex post, her participation constraint may not be met.

So the multiplicity of goals and principals implies that the provision of high-powered incentives are less likely to be suitable for the public sector than in the private sector where individuals may have to perform fewer, better defined, tasks. And, precisely because this is the case, the type of individuals found in the public sector may be more risk averse than those in the private sector. High-powered incentives may also not be necessary if it is the case that public sector workers are more motivated than those in the private sector by non-pecuniary benefits and other career concerns. (For more

[1] As an example, see Wilson's (1989) influential case study of US bureaucracy.

discussion of incentives, see the articles in this volume by Burgess and Ratto and by Besley and Ghatak).

The features of multiple principals and multiple tasks mean that the goals of a public agency may be in conflict. Consequently the performance measures used to evaluate often complex public sector performance may also be in conflict. The multiple and sometimes vague goals of the public sector mean that performance relative to these goals is difficult to measure. Individuals will respond to performance measures (PMs) in ways that maximise their own utility or benefit. This is not necessarily consistent with PMs improving welfare, nor is it necessarily in ways that are expected by those that design the system. The literature has many examples of both distorted measures and altered behaviour to improve the PM at the expense of unmeasured actions. Regular (usually annual) summary, outcome-based PMs can be altered, and they can change behaviour, possibly in dysfunctional ways. Examples abound in the public sector: they include massaging of truancy rates in UK education (Fitz-Gibbon 1996), massaging of waiting lists and treated cases in UK healthcare (see Smith 1995), unnecessary changes in the timing of graduation of workfare enrolees from schemes in the US (Courty and Marschke 1997). In economic terms, these measures can be (and are) 'gamed'. Gaming can take many forms. Smith (1995) has given a list of unintended consequences of publishing PMs in the public sector. These include tunnel vision; myopia; measure fixation; sub-optimisation; gaming; misrepresentation and misinterpretation[2]. While these are different forms of behaviour, all are due to the fact that the agent has different aims from the principal(s). As the principal tries to get higher effort (and so better public services) by implementing performance measurement, the response may be better services but also may be other less desired behaviour.

Finally, from a rather different perspective, Le Grand (1997) argues that the view of the motivations of those providing, funding and receiving welfare from the UK welfare state has changed. From its inception in the late 1940s to the mid-1970s providers and funders of welfare services were seen as 'knights', eschewing self-interest to achieve the collective good. The users of the service, in contrast, were seen as passive 'pawns' prepared to take what they were given without complaint. This view then changed to one where in which all parties were viewed as pursuing their own self-interests: in Le Grand's terms, they behaved as 'knaves'. Le Grand points out that in fact it is likely that individuals have a mixture of motivations and that design of the welfare state is better when it allows for this mixture of motivation. He also points out that the design of incentives may make individuals change their motivations. For example, he argues that giving high-powered financial rewards to doctors may turn them from knights to knaves, or at least increase the amount of knavish behaviour. In the context of performance management, this perspective emphasises the endogeneity of provider motivation to the type of performance management scheme. In other words, not only may individuals 'game' the system but the introduction of different methods of measuring performance and rewarding performance may attract different types of individuals to provide public services.

[2] Goddard, Mannion and Smith (2000) show how these can be derived from a principal-agent model.

3. What is performance management?

Rationale for a performance management system

Performance management may be undertaken at various levels of government and its purpose may differ depending on the level at which is it implemented. For example, it can be used to improve the performance of individual units (such as particular schools, hospitals, police forces). This may or may not be linked to 'best practice' exercises, in which the best performing units are used as an example for others to follow. It can be used as part of an attempt to improve the performance of the *overall* organisation. In this case the focus of the exercise is to improve the performance of the parent organisation as a whole, as well as possibly providing some developmental information for a single unit. For example, performance measurement may improve the overall performance of the education system even if it does not give many clues of itself to the problems within any one school. It may be used as part of the attempt to foster or generate pseudo-competition, for example, where purchasers in health care buy care from providers on the basis of measures of performance. It may also be used to improve accountability in the public sector (for example, to highlight 'failing schools'). It can be used as part of a resource allocation system, for example to enable central government to allocate funds to service providers such as local government agencies. In the use made of performance management in the UK public sector we can see elements of all these aims, but they are often not clearly separately identified and the same tools may be used for different purposes.

There are various possible ways in which performance measures may be used (Burgess et al 2002). The PM information may be kept internal to the organisation and not published. In this case it is a management tool. Alternatively, the PM may be made public. In this case it may be linked to an incentive scheme. If so, the scheme may be explicit or implicit. In an explicit scheme a direct financial reward is made available to either the individual, a subgroup of the organisation (if one can be defined) or the whole organisation. This is basically pay for performance at the organisational or sub-organisational level. The Public Service Agreements used by Treasury to give resources to government departments is an example of such a scheme. Under an implicit scheme, the organisation (and not the individual) gets a financial reward as a result of the response of others to the PM. A classic example of this is a quasi-market, in which providers of services are rewarded for good performance by getting more contracts. Even where there is no incentive scheme, explicit or implicit, publication of PMs may still have an effect on behaviour, for example through individuals' pride in their 'league position', or avoiding a label of being a 'failing' organisation. This is the idea behind 'name and shame' policies applied to schools. Conversely, there may be rewards to agents who perform well. These may be pecuniary or non-pecuniary. The published school league tables provide two types of non-pecuniary benefit to teachers, for example: a short run 'prestige' benefit as well as a longer run 'working conditions' benefit if good pupils are able to choose schools with a high league table position and if teachers prefer teaching good students[3]. In all these cases performance management is intended to

[3] Precisely how such mechanisms operate will depend on how capitation payments are calculated and on whether parents can effectively exercise school choice, on whether schools are capacity constrained, on whether they can select pupils, and so on. Here there is a trade-off between the efficiency properties

provide competitive pressure on organisations to improve, but the precise way in which it brings about better results differs.

To date, it is more common for PMs in the UK public sector to be linked to an implicit incentive scheme; one given in the form of client/service user/customer choice. The PMs then empower the client to make an informed choice. Such schemes were introduced in the UK as part of the quasi-market reforms in health, community care, housing and education. In all cases provider organisations were to get contracts on the basis of their performance. Initially, there were few measures of performance, but over time their number has increased. We review the PMs currently employed in health and education below.

Types of performance measure used

Performance measures come in a variety of forms. At one end of the spectrum, there are measures derived from an in-depth evaluation of an organisation's processes and outcomes, typically involving a site-visit and large amounts of documentation. UK examples are OFSTED visits of schools, police inspections, QAA of teaching in universities, HMI Prison reports. At the other end are measures that are derived from administrative data and comprise the collection and publication of summary performance measures. UK examples are the truancy and limited measures of national exam pass rates published for UK schools. In between there are measures based on the collection of detailed data collected specifically for the purpose of performance management, for example, so called 'report card' data that is collected on the quality of health care in the US. Report card data typically involves specific surveys of several aspects of the medical care of patients undergoing a treatment for a particular medical condition. The focus of this paper is primarily on the use of less detailed measures based on routine administrative data. In passing it should be noted that the more detailed measures are more expensive to collect, so if it can be shown that the summary measures provide as good a measure as more detailed ones, there is then a case for moving to such measures. Recent work in health in the US (Dranove et al 2002) shows that some summary measures may be as good as much more detailed expensive measures for one particular treatment (Acute Myocardial Infarction). There has been no comparable research in the UK.

It will seldom be possible – or optimal – for there to be one (scalar) performance measure. Both the designers of the performance management system and final users care about vectors of outcomes, but the weights these two principals attach to these outcomes may differ. This raises the question of whether (i) the designer's vector of outcomes should be aggregated in a way which reflects her marginal rate of substitution between the component outcomes or (ii) component scores should be made available so final users can apply their own weights based on their marginal rates of substitution. If a resource allocation mechanism is attached, then a scalar measure is required. If, however, users' preferences differ, then (ii) is better[4].

---

of the capitation system and how highly geared the system is in providing incentives. We are grateful to an anonymous referee for this point.

[4] We are grateful to an anonymous referee for making this point.

Moreover, while multi-dimensional PMs are harder to interpret, they are also more difficult to game.

Over time, there has been change in the form of PMs used in the UK public sector. In a review of the use of PMs across the public sector in the UK, Mannion and Goddard (2000) conclude that, across all the sectors they surveyed, there have been clear shifts in what data have been collected. These shifts are: from collection of data on a narrow range of dimensions of performance towards development of indicator packages which reflect a broader assessment of organisational activity; from gratuitous collection of performance data towards the development of more streamlined and focused indicator packages; and some development of cross-sector or interface indicators where it has been recognised that organisational performance is partly reliant on the actions of other agencies.

There has also been change in how PMs are used. Mannion and Goddard (2000) find there has been a general shift in the use of information on performance away from primarily being used for internal management control purposes towards use of these data for external accountability and control. Performance data has been increasingly used to mediate contractual relations. There has been a shift away from informal performance assessments based on peer review or sample-based inspection towards increased reliance on published performance league tables based on administrative data. More broadly, there has been a shift towards the use of performance information to facilitate participatory form of democracy and active citizenship. Finally, the current UK Labour administration has moved towards linking performance measures to financial rewards. Examples are the team based incentives schemes piloted in several government departments (Makinson 2000), or the greater access to capital investment and greater freedom from central management given to hospitals which perform well against a set of financial and other targets (Department of Health 2002).

Barnow (1992) identifies a range of performance measures which are used in government programmes. These measures do not necessarily correspond to any single economic concept and a programme may use more than one performance measure. The concepts he identifies include the following.

Gross outcomes

Gross outcomes are measures of outcomes of the programme at some designated date. Training programme gross outcomes include job placements, wages, exit from welfare. Health programme gross outcomes include the number of individuals who do not die after emergency admissions for heart attacks, or the number of individuals given hip replacements. Education gross outcomes include the number of pupils passing exams at a certain grade. The advantage of such measures are that they are easy to understand and easy to collect. These outcomes are gross in the sense that they do not necessarily measure the actual output of the programme itself. For example, pupils may have passed exams in the absence of school inputs, patients may have recovered from heart attacks without medical intervention. However, if gross outcomes are measured relative to some standard which is set to take account of what would have happened without the programme, then the gross outcomes may be useful in assessing the impact of the programme. Gross outcomes also do not take into

account the difficulty of treating a particular individual, but they can be adjusted for observed characteristics of the individual and this can partially overcome this problem. In fact, in many situations, heterogeneity across programme users means that such adjustments will be necessary. For example, performance measures for medical outcomes in the US are frequently risk adjusted (essentially adjusted for the health of the individuals treated). Under the JTPA programme, gross outcomes performance was adjusted for local labour market conditions (Barnow (1992) provides details).

Net outputs

These are measures of the value added of the programme[5]. In the JTPA context, value added can be thought of as the human capital added to the participants. As Barnow (1992) points out, barriers to developing net output measures may be high, primarily because the output of the programme may be multi-dimensional: for example, outputs from JTPA comprise a range of both general academic training and basic work skills. One way of overcoming this is to calculate the value of net outputs – the net earnings gain from participation – so there is a common metric across outcomes, but again this is not a simple matter. More generally, measuring value added is fraught with methodological problems, a key one being the difficulty of constructing a counterfactual of what would have happened in the absence of the programme. This means that net output calculations often require evaluations that are expensive and cannot be done on a regular basis. In addition, collecting such data takes time, so net measures do not give those who run the programme information when they need it, typically at the end of the year. In some sectors, e.g. education, net measures may be easier to collect. We discuss the use of value added measures of school performance later in this paper.

Inputs and processes measures

Examples of such measures include enrolees to programmes, number of staff, number waiting for treatment, teacher-pupil ratios. Whilst widely used, they provide no information on the effectiveness of the programmes. Using costs as a performance measure also biases activity towards shorter and less intensive programmes.

4. Assessing the performance of performance measures

Assessment of the impact of performance management is hampered by the lack of experimentation in, and associated assessment of, government policy. Performance measures have been introduced, generally not in a controlled trial manner, but as a result of a policy change. Experiments may be viewed as unethical or too expensive. Often performance assessment is accompanied by changes in other incentives. For example, the UK school league tables were introduced across all schools, and as part of the general reform of schooling provision. This makes it difficult to isolate the impact of introduction of performance management from other policy changes that are implemented at the same time. The long and short term impact of performance

[5] Note that value added here is not measured net of the cost of the programme. We return to this below in the context of value added PMs in education.

management may diverge, but the short term goals of politicians means that they not generally be interested in spending money undertaking evaluations which will only be known years after the implementation of a policy change.

The lessons from JTPA programme

The US Job Training Partnership Act (JTPA) provides one of the few instances in which an assessment of the impact of performance measures can be made. The JTPA introduced, in the early 1980s, a performance management system that was distinct in its focus on programme outcomes (e.g. job placements and trainee earnings) rather than outputs (e.g. persons trained), the use of budgetary incentives for managers based on outcomes, and the linking of performance measures across federal state and local government (Heinrich 2003). In addition, data was collected on the long-term earnings and labour market participation of enrolees to the programme. This has meant that the JTPA can be used to analyse the effectiveness of performance management systems based on outcomes.

The JTPA performance standards system mandates the provision of employment and training opportunities to "those who can benefit from, and are most in need of, such opportunities" (Heckman, Heinrich and Smith 1997 page 390). Basic performance standards are defined in terms of trainee employment, wage rates and earnings levels after the trainees leave the programme (at the time of most of the evaluations that have been undertaken of JTPA this was defined as the date at which the trainee was terminated from the programme). States implemented a variety of incentive schemes to reward good performance relative to the standards. Performance based incentives awards could only be spent to augment the budget of the training centres.

Heckman, Heinrich and Smith (1997, 2002) use JTPA data to examine the effectiveness of these performance measures (PMs). They ask three specific questions: first, do bureaucrats respond to incentives and are they equally effective for case workers and managers; second, do incentives point bureaucrats in the 'right direction'; and third, how much wasteful activity is induced by bureaucrats attempting to 'game' the standards?

They note first that the potential for conflict between efficiency and equity is written into the law authorising the programme. Second, PMs are written in terms of levels and not unobserved gains net of cost, the proper definition of efficiency. Third, PMs are, typically, short- rather than long-term measures. The use of these short-term outcomes creates the possibility that PMs misdirect activity by focusing training centre attention on criteria that may be perversely related to long term net benefits, long run equity criteria, or both. This is especially likely in a human capital programme, as one benefit of such a programme is that it encourages further education and training. Such additional investment depresses short run earnings but increases them in the long run.

The use of employment and earnings levels also gives rise to potential gains from cream-skimming. Training centres may select persons with high expected levels of target outcomes rather than those who would gain most from participation (have high

potential value added). Heckman et al (2002) define precisely the concept of cream skimming to assess whether it is a serious problem.

Their findings are as follows. First, they do not find cream-skimming to be such a problem[6]. They find instead the natural inclinations of employees – which are to help the most disadvantaged – often dominate, except where very performance orientated systems are run. They note even where this is not the case, the performance standards operate as a partial check on the preferences of case workers for helping the least advantaged (and perhaps those with the least value added). Second, they find that the long-term value added goals are not met. Instead, the short-term performance measures that are used in their place are either uncorrelated with, or are negatively correlated with, long-term value added. Third, they cite evidence from Courty and Marschke (1997) that managers game the system to achieve performance standards. Courty and Marschke (2004) find further evidence of such gaming behaviour and show that it has a negative impact on the true goal of the organisation.

Barnow (2000) also examines how closely the performance management criteria used to evaluate JTPA participants in 16 JTPA sites corresponds to the impact of the programme, where impact is defined from a randomised control study for those 16 JTPA sites (the National JTPA study). In undertaking this study he first notes that both the goals of JTPA have changed over time (e.g. in response to worries about 'cream-skimming') and the use of sanctions for failure have not always been imposed. Under the original design, poor performers, defined as those who failed to meet standards two years in a row, would be subject to reorganisation (which was essentially equal to loss of jobs). He notes that governors have been unwilling to do this, and instead performance measures have been modified. His findings echo those of Heckman, Heinrich and Smith (1997, 2002). He concludes that the current performance management system does not produce rankings that correspond closely to rankings on impact. For example, for several indicators/target group combinations, he finds that on the general criterion for sanctioning – falling in the bottom quarter of the performance rankings – a significant percentage of those ranked in the bottom quarter were misclassified. He also notes that there were difficulties in the performance management system controlling for local economic conditions.

There is also evidence from the JTPA programme that performance measures can be shaped by bureaucratic decisions. Heinrich (1999) studied one training centre (Cook County, Illinois) in which the performance incentives were strongly reinforced by the use of performance based contracts for the providers who delivered the services. Managers in Cook Country responded to the risk inherent in the reward system by passing the risks onto their service providers through the use of performance based contracts. The standards set in these contracts were higher than the ones the centre faced in its contracts. Overall, the strong performance incentives encouraged the selection of the more job-ready applicants and the provision of less intensive and less expensive training services. In contrast, in the Corpus Christi site studied by Heckman, Smith and Taber (1997) (and argued to be more typical) case workers maintained a strong preference for serving the most disadvantaged (and least employable).

[6] Cream skimming is defined precisely in terms of counterfactuals and is related to an economic model of performance standards in Heckman et al (2002). See also Heckman, Smith and Clements (1997).

In conclusion, Heckman, Heinrich and Smith (1997) note that it is important not to confuse a focused effort with a productive one. While managers responded to the performance management scheme by undertaking effort, and the gross outcomes based PMs improved as a result, this did not necessarily help to achieve the specific goals of the programme. Indeed, the goals themselves have evolved over time, through the influence of the bureaucrats within the programme.

The lessons from the education sector

The education sector in the UK has been subject to relatively high levels of public monitoring since the implementation of the 1988 Education Reform Act. Since then the performance management system has comprised two key elements: OFSTED reports and the annual publication of summary performance measures (PMs) in what are currently commonly referred to as the league tables. Here we focus on the latter. The aim of the publication of summary PMs is generally considered to be twofold: to create the incentive for schools to improve their students' educational attainment and to provide information on individual school performance to inform parental choice. The two are linked through the working of the quasi-market.

In the US the policy focus is on "improving the quality of schooling or how much is learned each year" for the same resource base (Hanushek 2002 page 2056). Emphasis has recently shifted from measurement of inputs to measurement of outcomes, with the publication of this information on report cards which sometimes include school rankings with respect to the various indicators. A variety of explicit and implicit incentives at both school and teacher level are incorporated into different states' accountability systems (Kane and Staiger 2002), with some options for exit from schools identified as failing. The use of accountability systems based on student test performance was put into law in January 2002 with the signing of the 'No Child Left Behind' Act (Cullen and Reback 2002).

So what are the outcome measures that try to capture movement towards the goal of improving student performance? Both in the UK and the US, one or more of the following are used, and in each case a school mean summary statistic is published (Kane and Staiger 2002). *Levels* are raw output scores (gross outcomes) of a cohort at a specific point in time, often reported as the percentage of that cohort achieving a particular target. In the UK, for example, one key PM is the percentage of a school's pupils who gain at least five GCSE passes at grade C or above. *Changes* aim to capture the improvement of successive cohorts at the same grade in the same school across time; while *gains* provide a measure of the progress of one cohort between two points in time. The value added measures published in the UK school performance tables for the first time in 2002 provide an example of a 'gain' PM (Wilson 2003).

The aim of using a value added PM is to better isolate the impact school environment has on pupil progress between two points in time. It does this by incorporating prior attainment, which helps to account for factors beyond the school's control, such as family background and other personal characteristics. In the UK, for example, two value added measures were published for each secondary school in the 2002 league tables: one provided an indicator of the average value added by the school between

the ages of 11 and 14 (Key Stage 2 and Key Stage 3) and the other between 14 and 16 (Key Stage 3 and GCSE – the latter being exams which mark the end of compulsory schooling)[7]. Details on how the latter UK value added PM is calculated are given in Wilson (2003). One point to note is that the UK value added PM takes no account of the resources used by a school. Similar to the net output PMs discussed by Barnow (1992), it therefore does not provide a measure of school effectiveness or efficiency, but rather a measure of total school performance, including teacher effects, resource levels and peer group (Meyer 1997). As such it may be suited more to the issue of parental choice rather than the government's aim of raising standards for the same resource base. A key point here is that a PM to reward schools should be measured net of peer effects, but a parent choosing a school should look at a value added PM which does not correct for them[8].

Koretz (2002) identifies three issues about inferring educators' performance from that of their students: (i) the limitations of measures employed; (ii) the perverse incentives that may be created; (iii) the difficulties in drawing inferences regarding the gains in student performance. Each type of PM raises distinct points with regard to these three issues; here we concentrate on levels and gains and draw on evidence from both the UK and the US.

Limitations of measures employed

The production of education is a complex process, so any one PM will at best be an imperfect measure of the multiple tasks undertaken by a school; indeed, some of these tasks may be inherently unmeasurable (Dixit 2002). Tests are inevitably incomplete samples of achievement domains (Koretz 2002) and hence may be subject to measurement error (Ladd and Walsh 2002). Kane and Staiger (2002) compare the statistical properties of the three types of PM. They show that PMs comprising raw output scores are subject to bias because they do not take account of factors outside a school's control[9]. While such levels-based PMs are the most reliable, much of this reliability is due to the unchanging characteristics of each school's population, rather than persistence in educational practices within the school. Incorporating prior attainment, i.e. using a value added PM, helps reduce the bias but at the cost of a reduction in reliability since the inclusion of prior attainment partially accounts for individual pupil characteristics. This leads to these PMs exhibiting a high degree of volatility year on year, which in turn reduces their predictive power. One solution may be to pool test score data over time: the NCLB Act gives states the option of using a three-year weighted average (Kane and Staiger 2002).

PMs based on raw test results provide a picture of the achievement of a group of pupils at a particular grade or level. But education is a cumulative process, building on a series of inputs over time (Hanushek 2002). Here again, value added represents an improvement on levels as it incorporates prior attainment and hence accounts for the impact of those inputs up to that point.

[7] The 2003 secondary school league tables will include a measure of 'whole school' value added, i.e. between the ages of 11 and 16.

[8] We are grateful to an anonymous referee for this point. See Hanushek (2002) for discussion of peer group effects.

[9] Goldstein (2003) discusses this point in relation to PMs that show changes across time.

While representing an improvement on levels based PMs, there are limitations to the value added PMs currently employed. Goldstein (2003) highlights three problems with regard to those published in the 2002 UK secondary school league tables. First, the input score used as the basis for the value added calculation is not sufficiently accurate: there is evidence that adjustment should be made for performance earlier in the child's school history. Second, the sometimes large uncertainty or confidence intervals for the resultant PMs should be emphasised, particularly for small school cohorts. These provide information on which schools are statistically different from the average and hence give context to the (in)accuracy of any subsequent rankings exercise. This is also an issue for levels based PMs (see also Kane and Staiger (2002) on this point with regard to US data). Third is the problem of mobility: students who are mobile tend to have different rates of progress and may be over-represented in certain schools; current value added PMs take no account of this.

Perverse incentives created

The possibility that the publication of summary performance measures may cause agencies to exhibit dysfunctional behaviour has been well documented following Smith (1995). Here we highlight three manifestations of such behaviour in the education context. The first follows from the fact that any summary PM will inevitably be an imperfect measure of a complex process. The incentive is therefore to concentrate on those parts of the process which are included in the summary measure, possibly to the detriment of other, less quantifiable, tasks. Wiggins and Tymms (2002) provide evidence of this 'narrowing' effect on the curriculum at primary level in the UK, Deere and Strayer (2001) show that passing rates on the tests included in the Texas school accountability system have increased relative to other tests, and Jacob (2002) shows how teachers responded to the test-based accountability system in Chicago public schools along several dimensions, one of which is substitution away from low-stakes subjects such as science and social studies.

Second, for the same quality of education received, the better the input (the higher the ability of the pupils) the better the output and hence the higher the school's relative position in a levels-based ranking exercise. A PM based on raw test scores does not explicitly account for heterogeneity in any population of students, so the school has the incentive to tailor the population to improve its indicator. There is evidence of schools undertaking various 'creaming' strategies in response to such PMs (Meyer 1997). Figlio and Getzler (2002) and Cullen and Reback (2002) both provide evidence of US schools reclassifying weak students in order that they are not eligible for the tests that are the subject of the indicator. There is anecdotal evidence from the UK that schools are removing weak students from GCSE courses and putting them into GNVQ equivalents (Times Educational Supplement 2002). Schools may additionally have the incentive to engage in cream skimming at the point of admission. Gerwitz et al (1995), Whitty et al (1998) and West and Pennell (2000) all discuss ways in which a UK school can design its admissions procedure in order that only certain types of pupils (and parents) are attracted to the school. These include the use of complicated admissions forms and pre-admission interviews. There is also indirect evidence from the UK that the publication of league tables (based on raw output PMs) creates the incentive to exclude certain types of pupils (Gillborn (1996) quoted in West and Pennell (2000)). One way to reduce the incentive for such behaviour is for the indicator itself to better account for such heterogeneity in the

pupil population, which is one argument supporting the use of value added measures of school performance.

Finally, different PMs give rise to specific incentives regarding how a school allocates its resources across any given pupil population. As stated above, levels based PMs are often reported in terms of the percentage of a school's pupils attaining a specific target. Such target indicators introduce an arbitrary dichotomy into continuous data and will therefore focus agents' attention on the borderline (Fitz-Gibbon and Tymms 2002). We may expect schools to shift their activities or target their resources to pupils who are expected to just miss the target in the absence of (extra) intervention. This may be to the detriment of pupils at either end of the ability distribution and may or may not be welfare improving. Wiggins and Tymms (2002) provide evidence of such behaviour in UK primary schools concerned with hitting their Key Stage 2 targets. Deere and Strayer (2001) show that it is those pupils at or below the passing level that exhibit the most improvement in the tests included in the Texas accountability system. While a PM based on a measure of value added should reduce the incentive for such dysfunctional behaviour, it should be noted that the current UK value added PM caps the output score at the eight best GCSE results or equivalent, which may create the incentive to distort effort away from the top end of the distribution (Wilson 2003).

Difficulties in drawing inferences regarding gains in student performance

Performance in the UK education sector – as measured by raw exam scores – has improved since the introduction of the quasi-market and the publication of the first league tables. It is, however, difficult to isolate which element(s) of such a huge programme of reform has had an impact on outcomes: there is no counterfactual to the introduction of performance management. Moreover, within the UK education sector, two types of performance measures have been introduced (summary PMs and in-depth OFSTED reports) and it is not clear to what extent one and/or the other has provided educators with the incentive to improve students' performance. Consequently there is a lack of evidence on the impact of performance management in education. Bradley at al (2000) provide evidence of the impact of the use of summary PMs in the UK. The main findings are that: (i) new admissions are positively related to a school's own exam performance and negatively related to the exam performance of its competitors in the same school district; (ii) the impact of the school's comparative exam performance on new admissions increased after the introduction of quasi-market forces; (iii) schools achieve better exam results when they are in competition with schools with good exam performance but the impact of this is small; (iv) excess demand for places in popular schools has led to an increase in capacity at those schools; (v) greater parental choice and increased competition have led to some polarisation with respect to family background. These results suggest that the publication of PMs (allied with implicit incentives via the quasi-market) has produced an improvement of outcomes, as measured by the PMs themselves. This is accompanied, however, by some evidence of a selection effect.

Evidence on the impact of accountability systems on student achievement in the US has been "limited and ambiguous" (Kane and Staiger 2002 page 106). There have been test scores increases in Texas and North Carolina, two states with highly visible accountability systems, but Kane and Staiger suggest caution in inferring causality for

three reasons. First, other states which were also engaged in performance measurement did not experience such increases in test scores. Second, at least part of the increase may have been due to both states' above average exclusion rates. Third, there is no evidence that the timing of improvement in performance in North Carolina schools coincided with the introduction of the performance measurement system. Clotfelter and Ladd (1996) reach a similar conclusion with regard to the introduction of the Dallas accountability system: Dallas test scores did rise relative to other Texas cities, but the timing of this improvement predated the accountability system by a year. Ladd (1999) extends this analysis of the impact of the Dallas program and finds some evidence of it having a positive impact on student outcomes: there are positive and relatively large effects for white and Hispanic seventh-graders, but not for their black peers. Jacob (2002) shows that test scores in maths and reading increased in Chicago public schools after the introduction of a test-based accountability policy, but that student effort and improvements in test-specific skills largely drove these improvements.

All the above evidence discusses and attempts to estimate the impact of performance management on measured outcomes. To date, these have been predominantly in the form of raw test scores or levels. An improvement in such outcomes does not necessarily represent an improvement in actual student learning, however, particular if their publication creates the perverse incentives discussed above (Kane and Staiger 2002). Judgement of the success of performance management in education in terms of the government's goals is hence particularly difficult. Moreover, there is additional difficulty in trying to attribute such improvements to teachers: much of the variance in test scores is controlled by non-school factors, and learning is a cumulative process rather than one which can be attributed to the impact of one teacher at any one point in time (Koretz 2002). Any inferences made about the incentives created by performance management must therefore be made with particular caution.

Recall that what is published in UK league tables (and on US report cards) is a summary statistic giving a measure of the average performance of the school with respect to each PM. In the 2002 UK secondary school league tables there are four PMs relating to GCSE exam scores, four for Key Stage 3 and two value added PMs, plus two improvement (or 'change') measures and indicators of absences, both authorised and unauthorised. While publication of multiple PMs may reduce the incentive for a school to game the more complex performance management system (Ladd 1999; Fitz-Gibbon 1997), there is obviously a trade off between complexity and transparency: it is increasingly difficult for parents to evaluate the information presented to them.

Two further issues are worth highlighting here. First, different PMs provide different rankings of relative school performance. Wilson (2003) shows just how sensitive a school's ranking position can be to the introduction of an alternative (value added) PM: in one case two schools moved from 8th and 9th (on the percentage of pupils gaining at least five GCSEs at grade C or above) to 21st and 22nd respectively – second-last and last place in their Local Education Authority's league table. Second, even if we accept that a value added PM provides a more accurate indication of the impact a school has on pupil progress, this is only reported at the level of school mean. Hence it gives information on the value added by a school to its average pupil. The use of such an aggregated measure may, however, hide differences in the value

added by the same school to pupils at different points in the ability distribution. Thomas (2001) and Wilson (2003) provide some evidence that there is such differential effectiveness across different schools. Hence it may be particularly difficult for parents to draw inferences on the potential impact of a school on their own child's future performance. Inaccuracy in such inferences will have a detrimental impact on the ability of parents to be effective drivers for improvement in the education market.

In summary, student outcomes as measured by raw test scores have improved since the introduction of performance management in both the UK and the US. Whether this represents an improvement in "the quality of schooling or how much is learned each year" (Hanushek 2002 page 2056) is more difficult to judge, as is to what extent the observed gains can be attributed to specific forms of performance measure rather than other, often concurrent, reforms.

The lessons from health care

The publication of data about performance in health care is not new, nor is it confined to the public sector. We focus on measures of quality and review evidence from both public and private health care systems, as many of the issues raised by the use of performance measurement are common to both.

The United States has the most recent experience of publication of data on health care performance (Marshall et al 2000). Information is available on the comparative performance of health insurance plans, hospitals and individual doctors. Many different organisations have contributed to this, including federal and state government, employers, consumer advocate groups, the media. Information is available about inputs, process measures and outcomes, though data on net outputs is rare. Perhaps the best known reporting system is the Health Plan Employer Data Information Set (HEDIS), which is produced by a not-for-profit partnership of private buyers, health plans and consumers. Information is provided at the level of plans, is based on both administrative and clinical data and covers both cost and quality. Participation is voluntary. Other high profile systems have focused on a single aspect of care, that of inpatient mortality (an outcome measure). The New York Cardiac Surgery Reporting system publishes data on hospital- and surgeon-specific risk adjusted coronary artery bypass surgery mortality. Similar work has been undertaken in other states. These data are adjusted for the type of patient treated and are derived from both administrative and clinical data.

In comparison, there have been few examples of release of information about quality of care in the UK (Marshall et al 2000). Some basic information about hospital performance in England and Wales has been available since the early 1980s. This was intended for use by hospital managers, but the consensus is that it was little used because it was not widely disseminated (and was also not linked to rewards (sanctions) for good (bad) performance). In Scotland, data comparing outcomes across both hospitals and health areas was made public from the early 1990s. Since 1999 there has been a large increase in the amount of publicly released data on the performance of English health care providers. The National Performance Assessment System measures the performance on a range of financial, clinical and patient care

indicators for hospitals, health care buyers and primary care providers. For acute hospitals there are currently (2003) 35 indicators of performance. These indicators have been released in terms of league tables for hospitals and NHS health care purchasers and will be released for primary care providers. For hospitals the 35 measures are currently amalgamated into a 4 level 'star system'. In the current amalgamation, greatest weight is given to financial and waiting list targets, and less weight to clinical performance and patient experience. From 2003 performance, as judged by this star system, will be linked to greater financial and managerial autonomy for the best performers and to replacement of management for those with no stars[10].

Performance measurement in health occurs in almost all OECD countries. In a review, Smith (2002b) argues that most nations have relied opportunistically on readily available data in the early stages of developing measures, but that most systems are now in the process of seeking out clinical outcome measures, patient satisfaction measures, and measures of population health in the 'hard to measure' aspects of health care performance.

Design issues

Smith (2002a) draws attention to a number of design issues that arise in the construction of performance measures in health care. The first is the need to adjust for sources of variation. Variation in outcomes may occur due to differences in case mix (the patients being served), resources used, priorities in outcomes, the external environment, accounting treatment, data errors and random fluctuations. For some purposes adjustment will need to be made for this variation: for example, when comparing morbidity from heart treatment, the clinical mix of patients needs to be taken into account. In other cases such adjustment is inappropriate: for example if a hospital has chosen to deliver services with an inappropriate mix of staff, adjustments mask this source of poor performance.

The second is the choice of process or outcome measures. While there has been a move towards outcome measures, there are cases in which process measures are useful. Smith (2002a) argues that outcomes are likely to be favoured when the nature of the outcome is relatively uncontested, when outcomes can relatively easily be captured in operational performance measures, when an indicator of the outcome can be secured reasonably soon after the intervention, when the outcome is readily attributable to clinical performance rather than external factors, and when there is a need for considerable clinical judgement as to the most appropriate intervention to offer. This list re-iterates some of the points earlier in this paper: outcome measures are of use in indicating good performance when the outcomes that are desired occur in the short term, when risk adjustments can be made, and when there are not too many dimensions to the output. Smith suggests this is more likely for acute than chronic treatments, as in the latter there may be no agreed consensus as to what constitutes a good outcome (for example, in much of primary care).

[10] See http://www.doh.gov.uk/performanceratings/2002/method_acute.html for more information on the 'star system'.

Process and outcome measures also might have different functions. Outcomes are often the result of factors outside the control of the health system. Measuring poor outcomes (e.g. inpatient mortality) does not necessarily give guidance on what to change in order to improve. Such outcomes occur infrequently in comparison to the processes that prevent them. On the other hand, changing a process that is known to improve outcomes should lead to improvements. So measuring processes may be useful when performance measures are used internally to improve production, whereas outcome measures may be more useful when seeking to reward good outcomes. Marshall et al (2000) argue that there is an increasing body of evidence that process measures are more sensitive and feasible measures of quality of care than outcome measures. More generally, a theme that emerges in the discussion of performance measures in health care is the importance of the use of a range of measures.

There is also discussion of the purpose of public disclosure of information. Marhsall et al (2000) identify three models. The first is a public accountability model in which disclosure is a public responsibility, independent of the consequences. Proponents argue that the release of data, in conjunction with appropriate education and debate, will clarify important social issues. However, this model may have little impact on the quality of care, though it may (for exactly the same reasons) be perceived as least threatening by professionals.

The second is a market orientated model, in which comparative information allows informed consumers to drive quality improvements through choice. To be useful, such data need to be standardised for differences in case-mix. This model is possibly more important in health care systems in which consumers can make choices, but even in the US, the evidence suggests that individual consumers make relatively little use of performance data.

The third model, a professional orientated model, assumes a desire on the part of professionals to improve their practice, given the appropriate environment. Providing data aids this. In this case, standardisation is not the highest priority, as providers may be comparing their own performance over time using the data as an aid, rather than a measure that is linked directly to financial rewards.

Evidence on the effectiveness of performance measures

Despite a decade or more of experience with public disclosure of performance data in the US there has been little rigorous evaluation of its impact. Marshall et al (2000) identify several reasons for this. Some of these are due to the particular nature of the sector, for example the 'political incorrectness' of challenging a tool of informed consumerism or the power of vested business interests. However, they also note that evaluation of performance data requires a clear theoretical framework to identify the purpose of publication and an understanding of the strengths and weaknesses of the data that are being made public. The kind of academic rigour that has been applied to evaluation of performance management in the training field has not yet occurred in health. In part, this is due to the fact that the measurement of healthcare outcomes is more complex than the measurement of labour market outcomes.

Given this caveat, Marshall et al identify possible responses to public disclosure from consumers, physicians, hospitals and provider organisations, and purchasers. The small amount of available evidence suggests that while consumers claim to want such information they do not make great use of the data (Schneider and Epstein 1998). Physicians are interested but sceptical about the data released. A survey of the impact of HEDIS data on employer choice of health care plans suggests limited use by buyers of health care (Hibbard et al 1997). There is some evidence that provider organisations and hospitals are most responsive to these data. Marshall et al (2000) indicate some positive responses, in that report cards induced better processes or outcomes, particularly in competitive markets. Other responses were less positive and included criticisms of the data.

One of the most fundamental questions is the extent to which disclosure influences the outcomes of care. The most reliable evidence comes from studies of the New York Cardiac Surgery Reporting System (Marshall et al 2000). Hannan et al (1994) found that mortality declined significantly following publication of data on mortality rates. Potential explanations include responses of doctors, an exodus of low volume and high mortality surgeons from the state, a marked improvement in the performance of non-low volume surgeons and improvement of surgeons new to the system. Critics of the release of data argued that its publication could have reduced access to CABG surgery by forcing sicker patients to get surgery out of the state or by surgeons refusing to operate on high risk patients (potential 'cream skimming' strategies). However, this behaviour was not supported by the evidence.

The recentness of published performance measures means there is little evaluation of the impact of such measures in the UK (as distinct from discussion). In his review of international evidence, Smith (2002b) notes both that there has been a large increase in the amount of data collected, and that attention is increasingly turning from the provision of data to its interpretation, often through to risk adjustment. However, he also notes that the weakest element of performance measurement systems is the response function, by which he means methods to ensure that management and clinical staff respond as intended to reported data.

In summary, the lessons from the experience in the health care sector appear to be the following. There are considerable technical difficulties in constructing measures of performance and value added (net output) measures seem very far from the current agenda. Second, several commentators argue the need for different measures for different purposes. It is suggested that some measures be used for internal use, others for external monitoring and regulation and yet others for giving information to consumers and other user groups. Third, the largest response to these measures, from the US literature, is by health care providers (both inside and outside the private sector). Finally, there has been little rigorous evaluation of the impact of performance measures.

5. Conclusion

Performance measures are now widely used within public sector organisations, but there is a lack of evidence regarding their usefulness. Hence it is still not clear to

what extent performance measures help agencies achieve the goals they have been set by policy makers. Several common themes emerge, however, from our review of the experience of performance management across the three sectors in the UK and the US.

First, there is some consensus that gross outcomes or levels based PMs do not provide a sufficiently accurate picture of the relative performance of public sector organisations, and that some adjustment should be made to account for the heterogeneity of the input. Such adjustment helps to (i) give a better measure of the impact of the agency and (ii) reduce the incentive for cream skimming. In the JTPA programme and the health care sector levels based PMs have been adjusted for local economic conditions and risk respectively. In education, the inclusion of prior attainment further improves the accuracy of the resulting value added, or net output, PM.

Second, a single PM is not sufficient. Public sector organisations often have multiple stakeholders who have differing, and sometimes conflicting, goals. One PM cannot adequately address all these actors' objectives. Instead a range of PMs should be employed, both in terms of what they measure and also in terms of their form.

Third, the intended purpose for each potential measure should dictate both its form and the decision whether to publicly disclose the resulting performance information. The use of PMs to improve process or performance within an organisation does not necessarily require the PMs to be published. If, however, the aim is to facilitate a market, then publication is required. In this case the accuracy of the information is essential in order that improvement in performance can be driven by effective competitive pressure.

The lessons we have drawn from our review of the use of performance measures in the public sector echo the more general public administration literature, which is concerned less with evaluation than with implementation of measures. For example, in a general discussion of effective performance measurement systems, Kravchuk and Schack (1996) suggest 10 design principles:

- Formulate a clear coherent mission
- Develop an explicit measurement strategy
- Involve key users in the design and development phase
- Rationalise the programme structure as a prelude to measurement
- Develop multiple sets of measures for multiple users
- Consider the customers throughout the process
- Provide each user with sufficient detail
- Periodically review and revise the measures
- Take account of complexities upstream, downstream and laterally
- Avoid excessive aggregation of information

Kravchuk and Schack also argue that these PMs should be used as indicators, not tools for management, which concurs with the current emphasis on public disclosure of performance information.

Public agencies exist precisely because there are conflicting goals amongst the stakeholders they represent. It is therefore unreasonable to expect that one set of

performance measures will solve the problem of governance in such bureaucracies. What is apparent from this review, however, is the need for more rigorous evaluation of the implementation of such schemes in order that their design may be improved in a systematic way. There is almost no evidence on the impact of public sector performance management schemes on outcomes, nor on the costs of achieving these outcomes. So there is almost no evidence on whether these schemes improve the efficiency of the public service being delivered. On the basis of our review, we would make the following recommendations.

First, piloting of performance management schemes should be considered more widely and, in particular, such pilots should be rigorously evaluated in order to provide the necessary evidence on their impact. Second, we need to better understand the link between process and outcome in order to ensure that monitoring of the former has the desired result on the latter. Finally, there may be scope for the development of targets based on alternative, independent information sources such as, for example, the British Crime Survey to set targets for police authorities, or the use of general household surveys to measure the health of people living in an area. These would be 'non-corruptible' indicators of performance; they are indicators that are not subject to manipulation by the individuals whose actions are being measured. Their use would force the relevant organisation to focus on what really mattered (for example, crime prevention, illness prevention) and it would also encourage them to find out what really mattered. Targets could be adjusted to take account of those differences between areas that are outside the control of those whose performance is being measured and hence the impact of public service providers on outcomes could be better isolated and compared. The use of such independent performance measures could therefore reduce the opportunity for dysfunctional behaviour, while maintaining the incentive to improve the efficiency of public service provision.

6. References

Barnow, BS (1992), The Effect of Performance Standards on State and Local Programs, in Manski , C and Garfinkel, L (eds) *Evaluating Welfare and Training Programme*, Cambridge MA, Harvard University Press

Barnow, BS (2000), Exploring the Relationship between Performance Management and Program Impact: A Case study of the Job Training Partnership Act, *Journal of Policy Analysis and Management*, 19(1): 118-141

Bradley, S, Crouchley, R, Millington, J and Taylor, J (2000), Testing for Quasi-Market Forces in Secondary Education, *Oxford Bulletin of Economics and Statistics*, 62(3): 357-390

Burgess, S, Propper, C and Wilson, D (2002), *Does Performance Monitoring Work? A Review of the Evidence from the UK Public Sector, Excluding Health Care*, CMPO, University of Bristol, Working Paper 02/049

Clotfelter, C and Ladd, HF (1996), Recognizing and Rewarding Success in Public Schools, in Ladd, HF (ed), *Holding Schools Accountable*, Washington DC, Brookings Institution: 23-63

Courty, P and Marschke, G (1997), Measuring Government Performance: Lessons from a Federal Job Training Programme, *American Economic Review Papers and Proceedings*, 87(12)(May): 383-388

Courty, P and Marschke, G (2004), An Empirical Investigation of Gaming Responses to Explicit Performance Incentives, *Journal of Labour Economics*, forthcoming

Cullen, JB and Reback, R (2002), *Tinkering Towards Accolades: School Gaming Under a Performance Accountability System*, University of Michigan (March)

De Bruijn, H (2002), Performance Measurement in the Public Sector: Strategies to Cope with the Risks of Performance Management, *The International Journal of Public Sector Management*, 15(7): 578-594

Deere, D and Strayer, W (2001), *Putting Schools to the Test: School Accountability, Incentives and Behaviour*, Department of Economics, Texas A&M University (March)

Department of Health (2002), Raising Standards Across the NHS, http://www.doh.gov.uk/raisingstandardsnhs/raisingstandardsnhs.pdf, 10/03/03

Dixit, A (2002), Incentives and Organizations in the Public Sector: An Interpretive Review, *Journal of Human Resources*, 37(4), 696-727.

Dranove, D, Kessler, D, McClellan, M and Sattherwaite, M (2002), *Is More Information Better? The Effects of 'Report Cards' on Health Care Providers*, NBER Working Paper 8697, NBER, Cambridge MA

Figlio, DN and Getzler, LS (2002), Accountability, Ability and Disability: Gaming the System?, http://bear.cba.ufl.edu/figlio/fig_getz.pdf (April), 14/10/02

Fitz-Gibbon, CT (1996), *Monitoring Education: Indicators, Quality and Effectiveness*, London, Cassell

Fitz-Gibbon, CT (1997), *The Value Added National Project: Final Report: Feasibility Studies for a National System of Value Added Indicators*, London, School Curriculum and Assessment Authority

Fitz-Gibbon, CT and Tymms, P (2002), Technical and Ethical Issues in Indicator Systems: Doing Things Right and Doing Wrong Things, *Education Policy Analysis Archives*, 10(6): http://epaa.asu.edu/epaa/v10n6, 12/02/02

Gerwitz, S, Ball, SJ and Bowe, R (1995), *Markets, Choice and Equity*, Milton Keynes, Open University Press

Gillborn, D (1996), *Exclusions from School*, Viewpoint Number 5, Institute of Education, University of London

Goddard, M, Mannion, R and Smith, P (2000), Enhancing Performance in Health Care: A Theoretical Perspective on Agency and the Role of Information, *Health Economics*, 9: 95-107

Goldstein, H (2003), *A Commentary on the Secondary School Value Added Performance Tables for 2002*, http://www.ioe.ac.uk/hgpersonal/value-added-commentary-jan03.htm, 27/02/03

Hannan, EL, Kilburn, H, Racz, M, Shields, E and Chassin, MR (1994), Improving the Outcomes of Coronary Artery Bypass Surgery in New York State, *JAMA*, 271: 761-766

Hanushek, EA (2002), Publicly Provided Education, in: Auerbach, AJ and Feldstein, M (eds), *Handbook of Public Economics Volume 4*, Amsterdam, Elsevier Science BV: 2045-2141

Heckman, J, Heinrich, C and Smith, J (1997), Assessing the Performance of Performance Standards in Public Bureaucracies, *American Economic Review, Papers and Proceedings*, 87: 389-395

Heckman, J, Heinrich, C and Smith, J (2002), The Performance of Performance Standards, *Journal of Human Resources*, 37(4): 778-811

Heckman, J, Smith, J and Clements, N (1997), Making the most out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts, *Review of Economic Studies*, 64(4): 487-535

Heckman, J, Smith, J and Taber, C (1997), What do Bureaucrats do? The Effects of Performance Standards and Bureaucratic Preferences on Acceptance into the JTPA

Programme, in: Kibecap, G (ed), *Advances in the Study of Entrepreneurship, Innovation and Economic Growth, Volume 7: Reinventing Government and the Problem of Bureaucracy*, Grenwich Conn, JAI Press: 191-218

Heinrich, CJ (1999), Do Government Bureaucrats Make Effective use of Performance Management Information? *Journal of Public Administration Research and Theory*, 9(3): 363-393

Heinrich, C (2003) Measuring Public Sector Performance and Effectiveness, forthcoming in Peters, BG and Pierre, J (eds), *Handbook of Public Administration*, Thousand Oaks, CA, Sage

Hibbard JH, Jewett, JJ, Legnini, MW and Tusler, M (1997), Choosing a Health Plan: Do Large Employers use the Data? *Health Affairs* 16: 172-180

Jacob, BA (2002), *Accountability, Incentives and Behaviour: the Impact of High-Stakes Testing in the Chicago Public Schools*, NBER Working Paper 8968, NBER, Cambridge MA

Kane, TJ and Staiger, DO (2002), The Promise and Pitfalls of Using Imprecise School Accountability Measures, *Journal of Economic Perspectives*, 16(4): 91-114

Koretz, DM (2002), Limitations in the Use of Achievement Tests as Measures of Educators' Productivity, *Journal of Human Resources*, 37(4): 752-777

Kravchuk, R and Schack, R (1996) Designing Effective Performance Measurement Sytems under the Government Performance and Results Act 1993, *Public Administration Review* 56(4): 348-358

Ladd, HF (1999), The Dallas School Accountability and Incentive Program: an Evaluation of its Impacts on Student Outcomes, *Economics of Education Review*, 18(1): 1-16

Ladd, HF and Walsh, RP (2002), Implementing Value-Added Measures of School Effectiveness: Getting the Incentives Right, *Economics of Education Review*, 21(1): 1-17

Le Grand, J (1997), Knights, Knaves or Pawns? Human Behaviour and Social Policy, *Journal of Social Policy*, 26(2): 149-169

Makinson, J (2000), *Incentives for Change. Rewarding Performance in National Government Networks*, Public Service Productivity Panel

Mannion, R and Goddard, M (2000), *The Impact of Performance Measurement in the NHS: Report 3: Performance Measurement Systems: A Cross-Sectoral Study*, Report prepared for the Department of Health, Centre for Health Economics, University of York

Marshall, M, Shekelle, P, Brook, R and Leatherman, S (2000), *Dying to Know: Public Release of Information about Quality of Health Care*, London, Nuffield Trust

Meyer, RH (1997), Value-Added Indicators of School Performance: a Primer, *Economics of Education Review*, 16(3): 283-301

Osbourne, D and Gaebler, T (1992), *Reinventing Government*, Lexington MA, Addison-Wesley

Osbourne S, Bovaird, T, Martin, S, Tricker, M, Waterson, P (1995), Performance Management and Accountability in Complex Public Programmes. *Financial Accountability and Management*, 11: 19-37

Schneider, EC and Epstein AM (1998), Use of Public Performance Reports, *JAMA* 279: 1638-1642

Sharifi, S and Bovaird, T (1995), The Financial Management Initiative in the UK Public Sector: The Symbolic Role of Performance Reporting, *International Journal of Public Adminstration*, 18: 467-90

Smith, P (1995), On the Unintended Consequences of Publishing Performance Data in the Public Sector, *International Journal of Public Administration*, 18 (2/3): 277-310

Smith P (2002a), *Some Principles of Performance Measurement and Performance Improvement*, Report Commissioned for Commission for Health Improvement, mimeo, University of York

Smith P (2002b), *Progress in Measuring the Health System Performance: Some International Experiences,* Report Commissioned for Commission for Health Improvement, mimeo, University of York

Thomas, S (2001), Dimensions of Secondary School Effectiveness: Comparative Analyses Across Regions, *School Effectiveness and School Improvement*, 12: 285-322

Times Educational Supplement (2002), *League Table Bonus Attracts Schools to Vocational Option*, 23 August

West, A and Pennell, H (2000), Publishing School Examination Results in England: Incentives and Consequences, *Educational Studies*, 26(4): 423-436

Whitty, G, Power, S and Halpin, D (1998), *Devolution and Choice in Education: the School, the State and the Market*, Milton Keynes, Open University Press

Wiggins, A and Tymms, P (2002), Dysfunctional Effects of League Tables: A Comparison Between English and Scottish Primary Schools, *Public Money and Management*, 22(1): 43-48

Wilson, JQ (1989), *Bureaucracy*, New York, Basic Books

Wilson, D (2003), *Which Ranking? The Use of Alternative Performance Indicators in the English Secondary Education Market*, CMPO, University of Bristol, Working Paper 03/058