ARTICLES

WHAT'S MEASURED IS WHAT MATTERS: TARGETS AND GAMING IN THE ENGLISH PUBLIC HEALTH CARE SYSTEM

GWYN BEVAN AND CHRISTOPHER HOOD

In the 2000s, governments in the UK, particularly in England, developed a system of governance of public services that combined targets with an element of terror. This has obvious parallels with the Soviet regime, which was initially successful but then collapsed. Assumptions underlying governance by targets represent synecdoche (taking a part to stand for a whole); and that problems of measurement and gaming do not matter. We examine the robustness of the regime of targets and terror to these assumptions using evidence from the English public health service on reported successes, problems of measurement, and gaming. Given this account, we consider the adequacy of current audit arrangements and ways of developing governance by targets in order to counter the problems we have identified.

MANAGING PUBLIC SERVICES BY TARGETS: AND TERROR?

In the mid-eighteenth century, Voltaire (in *Candide*) famously satirized the British style of naval administration with his quip 'ici on tue de temps en temps un amiral pour encourager les autres'. In the early twentieth century, the USSR's communist czars combined that hanging-the-admirals approach with a system of production targets for all state enterprises. The basic system survived for some 60 years, albeit with various detailed changes over time, before the Soviet system finally collapsed in 1991 (Ericson 1991) – a decline that has been attributed by some to not hanging enough admirals to counter gaming produced by the target system.

Gwyn Bevan is Professor of Management Science in the Department of Operational Research and LSE Health and Social Care, London School of Economics & Political Science. Christopher Hood is Gladstone Professor of Government and Fellow of All Souls College, University of Oxford.

Public Administration Vol. 84, No. 3, 2006 (517–538)

In the 2000s, Tony Blair's New Labour government in Britain adopted a watered down version of that system for performance management of public services, especially those in England. Having tagged a new set of government-wide performance targets onto the spending control system in 1998, in 2001 it added a key central monitoring unit working directly to the Prime Minister. From 2001, in England, the Department of Health introduced an annual system of publishing 'star ratings' for public health care organizations. This gave each unit a single summary score from about 50 kinds of targets: a small set of 'key targets' and a wider set of indicators in a 'balanced scorecard' (Secretary of State for Health 2001a, 2002a; Commission for Health Improvement 2003a, b; Healthcare Commission 2004). While the Blair government did not hang the admirals in a literal sense, English health care managers (whose life was perceived to be 'nasty, brutish and short' even before the advent of targets: Cole 2001) were exposed to increased risk of being sacked as a result of poor performance on measured indices (Shifrin 2001) and, through publication of star ratings, also to 'naming and shaming' (Anonymous 2001) (something that had been applied to schools and local government in the previous decade). Although there have been developments in performance assessment of public health care organizations in other UK countries following devolution, the policy context differed from England (Greer 2004): there was no emphasis on a few key targets, nor publication for 'naming and shaming'; nor was performance assessment linked with direct sanctions or rewards (Scottish Executive Health Department 2003; Farrar *et al.* 2004; Auditor General for Wales 2005). Hence these countries offer a natural experiment in assessing the impacts of the system of star ratings.

This paper seeks to explore some of the assumptions underlying the system of governance by targets and to expose those assumptions to a limited test based on such evidence as is available about responses to targets in the English public health care system up to 2004. How far did the system achieve the dramatic results associated with the Soviet target system in the 1930s and 1940s? Did it, for instance, produce a real breakthrough in cutting long waiting times – a chronic feature of the pre-targets system for 40 years – and how far did it produce the sort of chronic managerial gaming and problems with production quality that were later said to be endemic in the Soviet system? And to the extent that target systems of this type invite gaming by managers and other actors, are there ways of making targets and performance measures less vulnerable to gaming without scrapping them altogether?

THE THEORY OF GOVERNANCE BY TARGETS AND PERFORMANCE INDICATORS

Governance by targets and measured performance indicators is a form of indirect control necessary for the governance of any complex system (Beer 1966). The form of control that target systems represent is a version of homeostatic control in which: (1) desired results are specified in advance in

measurable form; (2) some system of monitoring measures performance against that specification; and (3) feedback mechanisms are linked to measured performance. Ironically perhaps, just as the targets system was collapsing in the USSR, the same basic approach came to be much advocated for public services in the West by those who believed in 'results-driven government' from the 1980s (see Pollitt 1986; Carter *et al.* 1995; Bird *et al.* 2005). It resonated with the ideas put forward by economists about the power of well-chosen *numéraires* linked with well-crafted incentive systems. It often appealed to public managers themselves as well because it could be portrayed as an alternative to the 'double-bind' approach to governing public services, one in which agents must strive to achieve conflicting and often not-fully-stated objectives, such that they fail whatever they do (Dunsire 1978). It also gave managers of complex, pluralistic, professional-heavy public organizations an explicit *rôle* and *raison d'être*.

Targets are sometimes kept secret. The type of regime considered here, however, is one in which targets and measures are published. Performance against those measures is also published (a principle going back at least to Jeremy Bentham's plans for prison management in the 1790s). The rewards and sanctions include: reputational effects (shame or glory accruing to managers on the basis of their reported performance); the award of bonuses and renewed tenure for managers that depend on performance against target; 'best to best' budgetary allocations that reflect measured performance; and the granting of 'earned autonomy' (ascertained from detailed inspection and oversight) to high performers. The last, a principle associated with Ayres and Braithwaite's (1992) idea of 'responsive regulation', was enshrined as a central plank in the New Labour vision of public management in its 1999 *Modernizing Government* White Paper (Cabinet Office 1999), as well as in a major review of public and private regulation at the end of its second term (Hampton 2004).

Such rewards and sanctions are easy to state baldly, but are often deeply problematic in practice. Summary dismissal of public managers can be difficult (as was the case even in the USSR in its later years). The 'best to best' principle of budgetary allocation will always have to confront rival principles, such as equal shares or even 'best to worst' (implying give the most to the weakest or most disadvantaged units) (Auditor General for Wales 2005). In addition, the earned autonomy principle of proportionate response implies a high degree of discretion accorded to regulators or central agencies that rubs up against rule-of-law ideas of rule-governed administration.

There are also major problems of credibility and commitment in any such system, given the incentives to 'cheat' both by target-setters and target managers (see Nove 1958; Miller 1992; Kornai 1994; Smith 1995; Heinrich 2002; Hood 2002; Propper and Wilson 2003; Bird *et al.* 2005). One possible way of limiting cheating and establishing commitment is by establishment of independent third parties as regulators or evaluators (Majone 1996; Power 1999). In the English variant of governance by targets

and performance indicators in the 2000s – in contrast to the Soviet model – semi-independent bodies of various types, often sector-specific, figured large in the institutional architecture alongside central agencies and government departments. But the commitment and credibility such bodies could add was precarious, given that most of them had only limited independence.

We now consider two linked assumptions that underlie the theory of governance by targets. One is that measurement problems are unimportant, that the part on which performance is measured can adequately represent performance on the whole, and that distribution of performance does not matter. The other is that this method of governance is not vulnerable to gaming by agents.

Assumptions about measurement: synecdoche

As indicated in figure 1, governance by targets implies the ability to set targets relating to some domain (small or large) of total performance which is to be given priority. That domain is here denoted as α, with performance outside that domain (β) assigned lesser importance. So the task is to develop targets measured by indicators, here denoted as M[α], to assess performance on α. The problem, as stated by Carter *et al.* (1995, p. 49), is that most indicators are 'tin openers rather than dials: by opening up a can of worms they do not give answers but prompt investigation and inquiry, and by themselves provide an incomplete and inaccurate picture'. Hence, typically, there will be a small set of indicators that are 'dials' – good measures (M[α_g]) for a subset of α, here denoted as α_g; a larger set of 'tin openers' – imperfect measures (M[α_i]) for another subset of α for which there are data available, here denoted as α_i, liable to generate false positives and/or false negatives; and another subset of α, here denoted as α_n, for which there are no usable data available. Accordingly, governance by targets rests on the assumptions

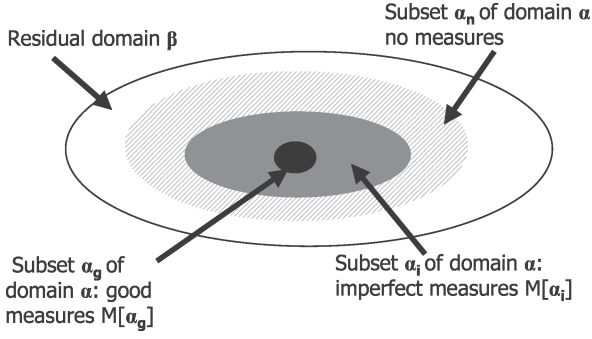

FIGURE 1 *Targeting priorities*

(i) that any omission of β and α_n does not matter; and
(ii) *either* that $M[\alpha_g]$ can be relied on as a basis for the performance regime, *or* that $(M[\alpha_g] + M[\alpha_i])$ will be an adequate basis for that regime.

What underlies these assumptions is the idea of synecdoche (taking a part to stand for a whole). Such assumptions would not be trivial even in a world where no gaming took place, but they become more problematic when gaming enters the picture.

Assumptions about gaming

Governance by targets rests on the assumption that targets change the behaviour of individuals and organizations, but that 'gaming' can be kept to some acceptably low level. 'Gaming' is here defined as reactive subversion such as 'hitting the target and missing the point' or reducing performance where targets do not apply (β and α_n). For instance, analysis of the failure of the UK government's reliance on money supply targets in the 1980s to control inflation led the economist Charles Goodhart to state his eponymous law: 'Any observed statistical regularity will tend to collapse once pressure is placed on it for control purposes' because actors will change their conduct when they know that the data they produce will be used to control them (Goodhart 1984, p. 94). And the 60-year history of Soviet targets shows that major gaming problems were endemic in that system.

Three well-documented gaming problems of the Soviet system were ratchet effects, threshold effects and output distortions. Ratchet effects refer to the tendency for central controllers to base next year's targets on last year's performance, meaning that managers who expect still to be in place in the next target period have a perverse incentive not to exceed targets even if they could easily do so (Litwack 1993): 'a wise director fulfils the plan 105 per cent, but never 125 per cent' (Nove 1958, p. 4). Such effects may also be linked to gaming around target-setting, to produce relatively undemanding targets, as James (2004, p. 410) claims to have applied to a number of Labour's public spending targets in the UK after 1998. Threshold effects refer to the effects of targets on the distribution of performance among a range of, and within, production units (Bird *et al.* 2005), putting pressure on those performing below the target level to do better, but also providing a perverse incentive for those doing better than the target to allow their performance to deteriorate to the standard (see figure 2), and more generally to crowd performance towards the target. Such effects can unintentionally penalize agents with exceptionally good performance but with a few failures, while rewarding those with mediocre performance crowded near the target range. Attempts to limit the threshold effect by basing future targets on past performance will tend to accentuate ratchet effects and attempts to limit ratchet effects by system-wide targets will tend to accentuate threshold effects. Attempts to achieve targets at the cost of significant but unmeasured aspects of performance (β and α_n) result in output distortions. Various such distortions were

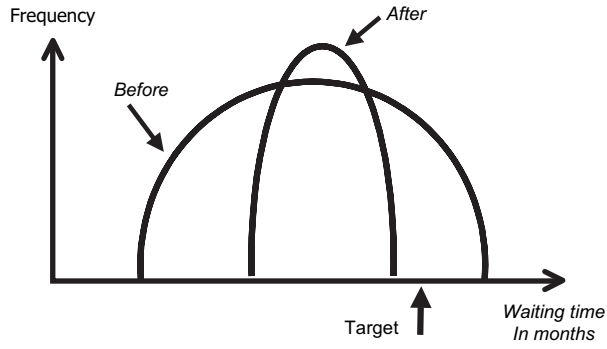

FIGURE 2 *Crowding towards the target*

well documented for the Soviet regime (Nove 1958, pp. 4–9), including neglect of quality, widely claimed to be an endemic problem from Stalin to Gorbachev (Berliner 1988, pp. 283–4).

The extent of gaming can be expected to depend on a mixture of motive and opportunity. Variations in the motives of producers or service providers can be described in various ways, of which a well-known current one is LeGrand's (2003) dichotomy of 'knights' and 'knaves'. Stretching that dichotomy slightly, we can distinguish the following four types of motivation among producers or service providers:

1. 'Saints' who may not share all of the goals of central controllers, but whose public service ethos is so high that they voluntarily disclose shortcomings to central authorities. A striking example of such behaviour in the English public health care system was exhibited in 2000 by St George's Healthcare NHS Trust, which twice drew attention to its own failures after two series of bad runs in its heart and lung transplantation programme and suspended its transplant work itself before its status as a designated centre was withdrawn by government (Commission for Health Improvement 2001, pp. 8–10).
2. 'Honest triers' who broadly share the goals of central controllers, do not voluntarily draw attention to their failures, but do not attempt to spin or fiddle data in their favour. Within the English public health care system, a notable example of 'honest trier' behaviour was exhibited in the 1990s by the Bristol Royal Infirmary, which did not attempt to conceal evidence of very high mortality in its paediatric cardiac surgery unit. The problem turned into a major scandal, but the official inquiry report into the issue began by saying that 'The story of paediatric cardiac surgical service in Bristol is not an account of bad people. Nor is it an account of people who did not care, nor of people who wilfully harmed patients' (Secretary of State for Health 2001b, p. 1).

3. 'Reactive gamers' who broadly share the goals of central controllers, but aim to game the target system if they have reasons and opportunities to do so. Such behaviour was highlighted by a question from a voter that apparently nonplussed Prime Minister Tony Blair during the 2005 British general election campaign – that a target for general practitioners in England to see their patients within 48 hours meant that in many cases primary care trusts would not book any appointments more than 48 hours in advance (Timmins 2005).
4. 'Rational maniacs' who do not share the goals of central controllers and aim to manipulate data to conceal their operations. In the English public health care system, a notorious example of a 'rational maniac' is that of the late Dr Harold Shipman who, as a general practitioner, killed at least 215 of his patients between 1975 and 1998 (Secretary of State for Health 2002b, Summary, paras 17–22). Shipman was a 'rational maniac' in that he appeared to be able to stop killing when he had good reason to think he was under suspicion (Secretary of State for Health 2002b, Chapter 13, paras 13.68–13.74). Although Shipman was (we hope) exceptional, Kinnel (2000) claims 'medicine has arguably thrown up more serial killers than all the other professions put together, with nursing a close second'.

Gaming as defined above will not come from service providers in categories (1) and (2) above (though there may be problems about measurement capacity as discussed in the previous sub-section at least for (2)), but will come from those in categories (3) and (4). Accordingly, governance by targets rests on the assumption that

(i) a substantial part of the service provider population comprises types (1) and (2) above, with types (3) and (4) forming a minority;

and

(ii) that the introduction of targets will not produce a significant shift in that population from types (1) and (2) to types (3) and (4)

or

(iii) that $M[\alpha_g]$ (as discussed in the previous sub-section) comprises a sufficiently large proportion of α that the absence of conditions (i) and (ii) above will not produce significant gaming effects.

These assumptions are demanding. LeGrand (2003, p. 103) argues that governance by targets can turn 'knights' into 'knaves' by rewarding those who produce the right numbers for target achievement, even if it means avoidance or evasion and neglect of β and α_n. Berliner (1988, pp. 289–90) observes that 'there have been heroic periods in the USSR when large numbers of people were selfless enough to provide the correct information required by planners to set taut but realistic targets [that is, functioned as actors of types (1) and (2)

above]', but argues that such periods were exceptional. Holmstrom and Milgrom (1991) in a classic model of how agents respond to incentives based on targets such as student performance in exams that omit key dimensions of performance (that is, where β and α_n are significant elements of performance), show that neither using a limited set of good signals ($M[\alpha_g]$) nor a larger set of poor signals ($M[\alpha_i]$) will produce results free from significant distortion by gaming. O'Neill (2002, pp. 43–59) argues similarly, albeit in different language, about performance assessment of professionals. So even if a target system begins with assumption (i) above being satisfied, a 'Gresham's law' of reactive gaming may mean that it fails to satisfy assumption (ii). (Gresham's law originally described the inevitability of bad money driving out good, but applied to governance by targets, it means that actors of types (1) and (2) above learn the costs of not gaming the system and shift towards type (3).)

If central controllers do not know how the population of producer units or service providers is distributed among types (1) to (4) above, they cannot distinguish between the following four outcomes if reported performance indicates targets have been met:

1. All is well; performance is exactly what central controllers would wish in all performance domains (α_g, α_i, α_n, β).
2. The organization is performing as central controllers would wish in domains α_g and/or α_i, but this outcome has been at the expense of unacceptably poor performance in the domains where performance is not measured (α_n, β).
3. Although performance as measured appears to be fine ($M[\alpha_g]$, $M[\alpha_i]$) actions are quite at variance with the substantive goals behind those targets (that is, 'hitting the target and missing the point').
4. There has been a failure to meet measured-performance targets ($M[\alpha_g]$, $M[\alpha_i]$), but this outcome has been concealed by strategic manipulation of data (exploiting definitional ambiguity in reporting of data or outright data fabrication).

In the section that follows, we consider how far the demanding assumptions identified here as underlying the theory of governance by targets were met in the English National Health Service under its 'targets and terror' regime of the early 2000s.

TARGETS AND TERROR AS APPLIED TO THE ENGLISH NATIONAL HEALTH SYSTEM (NHS)

The context and the institutional setting

The National Health Service (NHS) was created in 1948 as a UK-wide system for providing publicly organized and tax-financed health care for the population at large, replacing a previous patchwork system of regulated private, charitable and local authority organization. The organization that delivered the care was sub-divided into both functional units (acute hospitals) and

units defined territorially (care for the mentally ill, ambulances, primary care, dentistry, and so on), but broadly allowed clinical autonomy to medical professionals in their decisions on treating patients (Klein 1983; Hoque *et al.* 2004). Periodic reorganizations changed the boundaries, names and nature of those sub-units, but the system as a whole retained the features of block budgeting from central tax funds, public provision that was largely free (albeit with significant and growing exceptions for prescription drugs, dentistry and optical services), and the absence of any directly elected element in the organizational structure below the central ministry in London and its counterparts in Scotland, Wales and Northern Ireland. Observers of this health care system in cross-national comparative context such as Moran (1999) tended to see it as programmed to achieve (relative) cost containment at the expense of patient choice and some aspects of quality.

From the 1980s, there were various attempts to generate incentives for improved performance before the Blair government introduced its 'targets-and-terror' system for England in the early 2000s (Bevan and Robinson 2005). In the 1980s there were attempts to make hospital managers more powerful relative to medical professionals. In the 1990s a Conservative government introduced an 'internal market' into the public health care system in which providers were intended to compete with one another (Secretaries of State for Health for Health, Wales, Northern Ireland and Scotland 1989; Bevan and Robinson 2005). However, this system did not change the three basic institutional features described above and central government ministers continued to intervene to avoid hospitals being destabilized in the market (Tuohy 1999). In adapting this system after it won government in 1997, Labour tried to devise a control system that did not rely on funds moving between competing providers. Central to that new approach was the targets-and-terror system of governance of annual performance (star) ratings of NHS organizations that was referred to earlier.

By the mid-2000s this system applied to about 600 NHS organizations in England, comprising five different types of trust, and was part of a broader control system for public service performance. There were two central agencies: the Prime Minister's Delivery Unit which from 2001 monitored a set of key public-service targets for the PM by a 'war room' approach, of which two or three applied to health; and the Treasury, which from 1998 attached performance targets (Public Service Agreements or PSAs) to financial allocations to spending departments (James 2004), of which 10 or so applied to health care. In addition, the Department of Health continued to act as the overall overseer of the health care system, though operating increasingly at arm's-length from health care providers. There were also free-standing regulators of health care standards of which the main one (know as the Healthcare Commission at the time of writing) was responsible for inspections and performance assessment, including the published star ratings. Finally, there were two national audit organizations, the National Audit Office (NAO) that audited central government expenditure across the UK,

including the Department of Health's spending, and the Audit Commission, responsible for auditing the probity of NHS spending in England, as well as numerous other regulators and assessors of parts or all of the health care system. Walshe (2003, p. 153), for example, identified nearly 20 additional organizations of this kind (the numerous medical and surgical Royal Colleges are classed as one organization). Taken together, what lay behind the system of governance by targets in health care in the early 2000s amounted to an institutionally complex and frequently changing set of overseers, inspectors and assessors.

REPORTED PERFORMANCE DATA SHOWING IMPRESSIVE IMPROVEMENTS

On the face of it, the targets and terror system overseen by this army of monitors and assessors produced some notable improvements in reported performance by the English NHS. Three 'before' and 'after' comparisons in England and a fourth cross-country comparison relative to trusts elsewhere in the other UK countries without star ratings target systems may serve to demonstrate the point.

Figure 3 shows percentages of patients seen within the 4-hour target by the four quarters of each year in hospital Accident and Emergency (A&E) Departments (National Audit Office 2004). The star ratings required increases in this percentage each year from 2000–01. The National Audit Office (2004, p. 2) found that: 'Since 2002, all trusts have reduced the time patients spend in A&E, reversing a previously reported decline in performance. In 2002, 23 per cent of patients spent over four hours in A&E departments, but in the three months from April to June 2004 only 5.3 per cent stayed that long'. This reduction was achieved despite increasing use of A&E services, and the NAO also found evidence that reducing the time spent in A&E had increased patient satisfaction.

Figure 4 shows by ambulance trust the percentage of category A calls seen within 8 minutes for 1999–2000 and 2002–03 (Department of Health 2005). Category A calls are immediately life-threatening emergencies. The target of

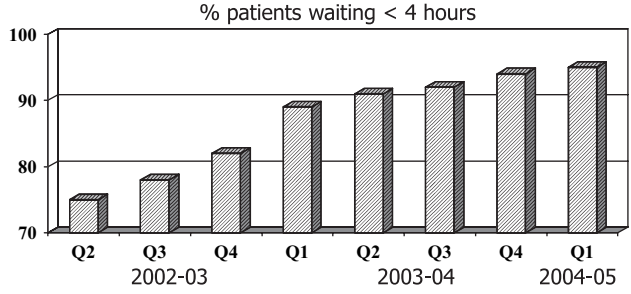

FIGURE 3 *Percentages of patients spending less tha 4 hours in A&E*

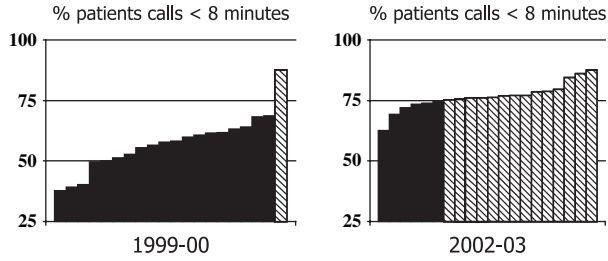

FIGURE 4 *Percentages of category A calls met within 8 minutes*

reaching 75 per cent of these within 8 minutes had existed since 1996. For 1999–2000, prior to star rating, some trusts only managed 40 per cent. After achieving 75 per cent became a key target for ambulance trust star ratings from 2002-03, performance jumped dramatically, and, at the end of that year, the worst achieved nearly 70 per cent.

Figure 5 gives numbers of patients waiting for first elective admission for more than 9 and 12 months at the end of March from 1997 to 2004 (Department of Health 2004). Maximum waiting times were dramatically reduced in England after the introduction of the star rating system from 2000–01. This set targets for maximum waiting times for the end of March each year; and for 2003 and 2004 these were 12 and 9 months.

Figure 6 gives percentages of patients on waiting lists waiting for first elective admission in each UK countries at the end of March from 2000 to 2003 (Office of National Statistics 2004). There was a notable difference between the dramatic improvement in reported waiting times for England, as against the other countries in the UK, which did not apply the targets-and-terror system of star ratings described earlier. Reported performance in the other countries did not in general improve, and at the end March of 2003, when virtually no patient in England was reported as waiting more than 12

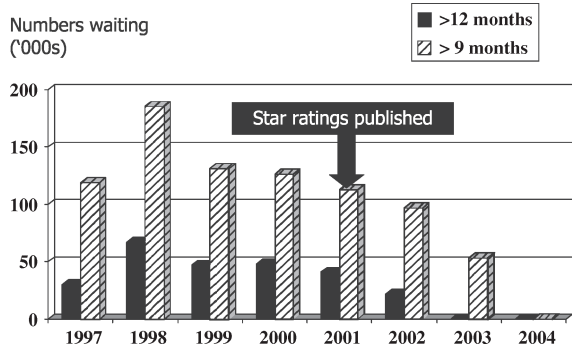

FIGURE 5 *Numbers waiting for elective admission in England*

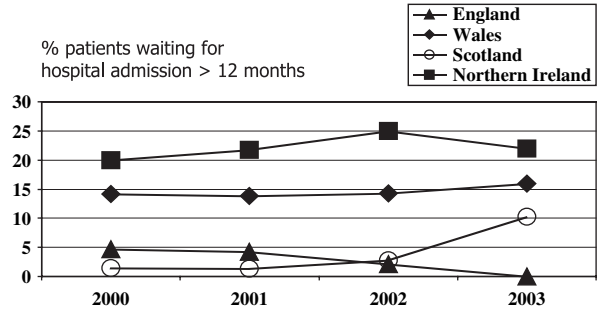

FIGURE 6 *Percentages of patients waiting more than 12 months for elective admission*

months for an elective admission, the equivalent figures for Scotland, Wales and Northern Ireland were 10, 16 and 22 per cent of patients respectively.

These improvements in reported performance are dramatic and on the face of it indicate the sort of results that the USSR achieved with its targets system from the 1930s to the 1960s, when it successfully industrialized a backward economy against a background of slump and unemployment in the capitalist West, emerged the victor in World War II and rebuilt its economy afterwards, to the point where, in 1961, its leaders publicly challenged the USA to an economic race over per capita production (Nove 1961, pp. 295–7). We now examine how far the control system met the assumptions we set out in the previous section.

THE ASSUMPTIONS REVISITED: MEASUREMENT AND GAMING

Measurement

On pages 520–1, above, we argued that governance by targets rests on the assumption (1) that the omission of β (and α_n if applicable) from performance measurement does not matter; and (2) *either* that $M[\alpha_g]$ can be relied on as a basis for the performance regime, *or* that ($M[\alpha_g] + M[\alpha_i]$) will be an adequate basis for that regime. In the case of health care these distinctions turn out to be central to the design of any performance management regime.

At first sight, waiting times for access to care at first sight may appear to be a clear case of $M[\alpha_g]$, but even for this indicator several inquiries have revealed data limitations that are far from trivial. For A&E targets, the National Audit Office (2004) found weaknesses in arrangements for recording time spent and observed that the relevant management information systems mostly pre-dated the targets regime and some were over ten years old. There were apparent discrepancies between officially reported levels of performance and independent surveys of patients in achieving the target for patients spending fewer than four hours in A&E: in 2002/03, officially, in 139 out of 158 acute trusts 90 per cent of patients were seen in less than four hours, but only 69 per cent of patients reported that experience in the survey

(Commission for Health Improvement 2004); in 2004/05, the official level had increased to 96 per cent (Anonymous 2005), but the survey-reported level was only 77 per cent (Healthcare Commission 2005a). For ambulance targets, there were problems in the definition of what constituted a 'life-threatening emergency' (the proportion of emergency calls logged as Category A ranged from fewer than 10 per cent to over 50 per cent across ambulance trusts) and ambiguity in the time when the clock started (Public Administration Select Committee 2003, p. 18; Bird *et al.* 2005). For hospital waiting time targets, the Audit Commission (2003), on the basis of 'spot checks' at 41 trusts between June and November 2002, found reporting errors in at least one indicator in 19 of those trusts. As we shall stress later, there was no systematic audit of measures on which performance data are based, so such inquiries were both partial and episodic. But they raise serious questions as to how robust even the $M[\alpha_g]$ measure was for this performance regime – an issue to which we return in the section that follows.

As noted earlier, the quality problem bedevilled the Soviet targets regime and quality remained in the subset of α_n. Likewise, Pollitt (1986, p. 162) criticized the 1980s generation of health care performance indicators in the UK for their failure to capture quality in the sense of impact or outcome. And that problem had by no means disappeared in the 2000s targets-and-terror regime for health care governance in England. Methodologically, measures of effectiveness remained difficult, required new kinds of data that both were costly and problematic to collect, and tended to rely on indicators of failure (Rutstein *et al.* 1976). The star ratings of the 2000s, like the predecessor performance indicators of the 1980s failed to capture key dimensions of effectiveness. There was a large domain of unmeasured performance (α_n) and measures of 'sentinel events' indicating quality failures (notably crude mortality rates and readmission rates for hospitals) were at best indicators of the $M[\alpha_i]$ 'tin-opener' type (Bird *et al.* 2005). Risk-adjusted mortality rates could be calculated for a few procedures such as adult cardiac surgery. But even there, problems in collecting the detailed data required led to a failure to achieve a high-profile ministerial commitment – announced after the Bristol paediatric cardiac surgery scandal referred to earlier – to publish, from 2004, 'robust, rigorous and risk-adjusted data' of mortality rates (Carlisle 2004).

Gaming

On pages 522–3, we argued that governance by targets rests on the assumption that

(i) a substantial part of the service provider population comprises 'saints' or 'honest triers', with 'reactive gamers' and 'rational maniacs' forming a minority;

and

(ii) that the introduction of targets will not produce a significant shift in that population from the first to the second pair of categories

or

(iii) that M[α_g] (as discussed in the previous sub-section) comprises a sufficiently large proportion of α that the absence of conditions (i) and (ii) above will not produce significant gaming effects.

As mentioned above, there was no systematic audit of the extent to which the reported successes in English health care performance noted on pages 526–8, above, were undermined by gaming and measurement problems, even though much of the data came from the institutions who were rated on the basis of the information they provided. That 'audit hole' can itself be interpreted by those with a suspicious mind (or a long memory) as a product of a 'Nelson's eye' game in which those at the centre of government do not look for evidence of gaming or measurement problems which might call reported performance successes into question. In the Soviet system, as all bodies responsible for supervising enterprises were interested in the same success indicators, the supervisors, rather than acting to check, connived at, or even encouraged, gaming (Nove 1958, p. 9; Berliner 1988, p. 37). In the English NHS, 'hard looks' to detect gaming in reported performance data were at best limited. Central monitoring units did mount some statistical checks on completeness and consistency of reported data, but evidence of gaming was largely serendipitous and haphazard, emerging from particular inquiry reports or anecdotal sources. We therefore cannot provide any accurate estimate of the distribution of the health care provider population among the four categories identified above (though examples of the existence of each of those types can be readily given, as we showed earlier). But even if we have to return a Scottish 'not-proven' verdict on assumption (i) above (that is, the evidence is insufficient either to accept or reject the validity of that assumption), assumption (ii) seems unsafe for the case being considered here, and, contrary to assumption (iii), there is enough evidence of significant gaming to indicate that the problem was far from trivial.

On pages 521–2, above, we discussed three main types of gaming identified in the literature on targets and performance indicators, namely ratchet effects, threshold effects and opportunistic output distortions. Here we concentrate on the third type of gaming, although there is some evidence of the presence of the first two types as well. Goddard *et al.* (2000) found clear ratchet effects in health care cost targets in the 1990s. As for threshold effects, figure 4, above, shows that ambulance trusts sought to meet the 75 per cent response-time target but not exceed it, and there were strong allegations that some ambulance trusts achieved this result by relocating depots from rural to urban areas. Insofar as this strategy meant that those who lived in rural areas would wait longer than the 8-minute target, it meant that the aggregate target could not be far exceeded (Commission for Health Improvement 2003c).

We now present evidence of gaming through distortion of reported output for ambulance response-time targets, hospital A&E waiting-time targets and hospital waiting time targets for first outpatient appointment and elective admission. A study by the Commission for Health Improvement (2003c) found evidence that in a third of ambulance trusts, response times had been 'corrected' to be reported to be less than eight minutes. The kinds of different patterns discovered are illustrated by figure 7: an expected pattern of 'noisy decline' (where there has been no 'correction'), and of a 'corrected' pattern with a curious 'spike' at 8 minutes – with the strong implication that times between 8 and 9 minutes have been reclassified to be less than 8 minutes. There was also evidence that the idiosyncracies of the rules about Category A classification led in some instances to patients in urgent need being given a lower priority for ambulance response than less serious cases that happened to be graded Category A.

For hospital A&E waiting-time targets, five types of output-distorting gaming response were documented. First, a study of the distribution of waiting times in A&E found frequency peaked at the four-hour target (Locker and Mason 2005) – although this pattern was much less dramatic than that for ambulance response times. Surveys by the British Medical Association reported widespread practice of a second and third type of gaming responses: the drafting in of extra staff and the cancelling of operations scheduled for the period over which performance was measured (Mayor 2003, p. 1054; British Medical Association 2005). A fourth practice was to require patients to wait in queues of ambulances outside A&E Departments until the hospital in question was confident that that patient could be seen within four hours (Commission for Health Improvement 2003c). Such tactics may have unintendedly caused delays in responding to seriously ill individuals when available ambulances were waiting outside A&E to offload patients (for an example of a fatal case, see Howarth 2004). A fifth gaming response was observed in response to the so-called 'trolley-wait' target that a patient must be admitted to a hospital bed within 12 hours of emergency admission.

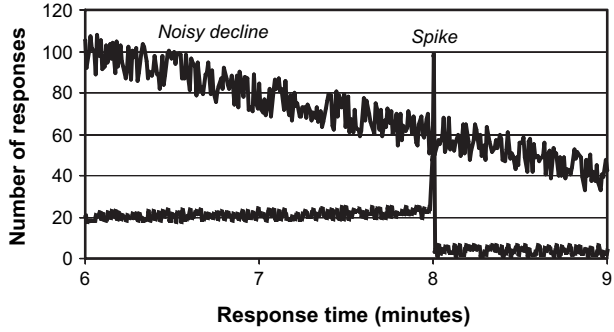

FIGURE 7 *Frequency distributions of ambulance response times*

The response took the form of turning 'trolleys' into 'beds' by putting them into hallways (Commission for Health Improvement 2002, para 3.19).

For hospital waiting time targets for first outpatient appointment and elective admission, the National Audit Office (2001) reported evidence that nine NHS trusts had 'inappropriately' adjusted their waiting lists, three of them for some three years or more, affecting nearly 6000 patient records. In five cases the adjustments only came to light following pressure from outsiders, though in four cases they were identified by the trusts concerned. The adjustments varied significantly in their seriousness, ranging from those made by junior staff following established, but incorrect, procedures through to what appears to be deliberate manipulation or misstatement of the figures. The NAO study was followed up by the Audit Commission, which, in its 2002 spot check study of 41 trusts referred to above, found evidence of deliberate misreporting of waiting list information at three trusts (Audit Commission 2003). In addition, a parliamentary select committee report on targets in 2003 reported that the waiting time target for new ophthalmology outpatient appointments at a major acute hospital had been achieved by cancellation and delay of follow-up appointments, which did not figure in the target regime. Recording of clinical incident forms for all patients showed that, as a consequence, 25 patients lost their vision over two years, and this figure is likely to be an underestimate (Public Administration Select Committee 2003, para 52).

Further, the publication of mortality data as an indicator of quality of clinical care may itself have produced reactive gaming responses. There is anecdotal evidence that such publication results in a reluctance by surgeons to operate on high risk cases, those who stand to gain most from surgery (Marshall *et al.* 2000). Because mortality rates are extremely low (about 2 per cent), one extra death has a dramatic impact on a surgeon's performance in a year, and risk-adjustment methods cannot resolve such problems.

These data, limited as they are, suggest that, relative to assumption (i), reactive gaming seems to have been practised by a significant minority of service-provider units (ranging from 7 to 33 per cent in the studies quoted), and that, relative to assumption (ii), star-rating-related targets seem to have produced an increasing share of organizations in the 'reactive gaming' category. Moreover, they suggest some significant problems about assumption (iii) that $M[\alpha_g]$ forms a large enough proportion of α to be proof against gaming effects. As the last example shows, synecdoche (taking a part for the whole) in target systems can be shown to have produced some clear negative effects on performance in the realms of β and α_n – the classic problem of the Soviet target system. Indeed, the star rating system meant that it was possible for three-star trusts to have within them a scandalously poor clinical service, and zero-star trusts an excellent service. Rowan *et al.* (2004) found no relationship between performance in star ratings and the clinical quality of adult critical care provided by hospitals. And indeed, in the examples of types of players given on pages 522–3, above, none of the quality failures at

Bristol, St George's and with Harold Shipman would have damaged the star ratings of the institutions concerned, because the types of mortality involved were relegated to the ß (or at best α_n) category.

DISCUSSION AND CONCLUSION

We have argued that the implicit theory of governance by targets requires two sets of heroic assumptions to be satisfied: of robust synecdoche, and game-proof design. And we have shown that there is enough evidence from the relatively short period of its functioning to date to suggest that these assumptions are not justified. The transparency of the system in real time seems to have exacerbated what we earlier described as Gresham's law of reactive gaming,

We see the system of star rating as a process of 'learning by doing' in which government chose to ignore the problems we have identified. A consequence was that although there were indeed dramatic improvements in reported performance, we do not know the extent to which these were genuine or offset by gaming that resulted in reductions in performance that was not captured by targets. Evidence of gaming naturally led many critics of New Labour's targets-and-terror regime to advocate the wholesale abandonment of that system. But the practical alternatives to such a regime (such as specific grants to providers to incentivize particular activities, true 'command and control' from the centre in terms of orders of the day, or governance by a double-bind approach that swings between unacknowledged contradictions) are well-tried and far from problem-free. Nor is health care truly governed by anything approximating a free market in any developed state: regulation and public funding (even in the form of tax expenditures) take centre stage in every case.

We conclude by considering how the theory and practice of governance by targets could be redesigned so that it is less vulnerable to gaming. Although gaming proved to be endemic in the much longer-lived Soviet targets regime, the prospects for a more game-proof design may be better in a mixed-economy system for delivering public services. Accordingly, we make suggestions for making systems of governance by targets more proof against synecdoche and gaming difficulties, by modified ways of specifying targets, measuring performance and monitoring behaviour.

Complete specification of targets and how performance will be measured almost invites reactive gaming by managers of service-providing units. Hence an obvious remedy is to introduce more uncertainty into these specifications (Bevan and Hood 2004) by making them transparent in process and in retrospect but not in real time. Such a design would follow Heald's (2003, p. 730) distinction between 'event' transparency and 'process' transparency, with 'assurance that established procedures have been followed and that relevant documentation is then placed in the public domain' (Heald 2003, p. 71). When targets take the form of general standards (as was

proposed for assessment by the Healthcare Commission (2005b) at the time of writing), advance warning of when assessments will be made will be of only limited value to potential gamers. But when targets for performance assessment are defined at a high level of specificity, there needs to be some uncertainty about the monitoring process. In the case of speed cameras, for example, drivers may know the cameras' locations from website or other sources, but do not know whether any particular camera is operating or what precise speed trips the camera into action. It is possible for a lottery to be fully transparent in a real-time process sense if the programming principles behind it can be fully revealed to the players, even if that does not enable them to know the actual numbers it will reveal. Introducing randomness into monitoring and evaluation in order to limit gaming violates only a very extended version of the transparency principle and one that is arguably not appropriate for performance monitoring.

Another way of limiting gaming would be to fill the 'audit hole' referred to earlier. Although British public services in general, and the English health care system in particular, groan under regulation and audit from various inspectors and auditors, audit of the data on which performance assessments are based is both fragmentary and episodic. As the existence of gaming becomes more generally recognized, failure to fill this hole invites the cynical view of the target regime as a 'Nelson's eye' game, in which central government colludes with those who game targets, by seeking improvements in reported performance only, and not providing the organizational clout to ask awkward questions about the robustness of those reported improvements. What is required is a new approach to performance data provision and auditing, similar to that of the 'Office of Performance Data' advocated by Robert Behn (2001).

A second means of monitoring would be by supplementing the arcane and impersonal process of reporting from one bureaucracy to another in a closed professional world by a greater face-to-face element in the overall control system. After all, in democratic theory the ideal of transparency is often seen as face-to-face communication between governors and governed, and even in the Soviet system it has been shown that public criticism of gaming by managers through the media was a salient feature of the overall system that served to limit managerial gaming. Indeed, it could be argued that face-to-face scrutiny of that kind is likely to be far less vulnerable to the gaming strategies that can undermine the target systems described here.

Of course, face-to-face interactions between health care providers and the public are far from problem-free (something graphically brought out by the Shipman case referred to on page 000, lines 00–00, above), and it is problems of that kind that has led to the targeting systems monitored by professionals. However, finding a way that an individual like Shipman will stand out from the vast majority (it must be hoped) of medical practitioners who are not serial killers requires, even in retrospect, elaborate statistical analysis. The final report of the Shipman Inquiry (Secretary of State for Health 2004) recommended using

a method of statistical monitoring of deaths in general practices which, using historical data, would have identified Shipman in 1988 (Aylin 2003). If such monitoring, using transparent thresholds, had been applied to Shipman when he was in practice, however, then it is likely that he would have managed his murder count and other deaths so that he would have avoided generating a statistical signal. Goodhart's law means that we may be able to use statistical analysis on historical data to generate a reliable signal when the people who generated the data knew that it would not be used for that purpose. But once the individuals concerned know the data they produce will be used for that purpose, their behaviour is likely to alter. Accordingly, if a transparent monitoring system were introduced in response to Shipman, this would probably fail to detect another rational maniac of the Shipman type, but put many other innocent GPs under suspicion of murder (Secretary of State for Health 2003).

Indeed, such a conclusion suggests that even and perhaps especially for the professional monitors, some face-to-face scrutiny mixed with random visitations may serve to limit the problems of synecdoche and gaming, particularly for organizations as complex as acute hospitals, given both ambiguity in definitions and noisy data. Since the 1990s in the US, the Joint Commission on the Accreditation of Health Care Organizations has been seeking to move towards a continuous process of monitoring hospital performance through performance indicators, but the foundation of its accreditation programme continues to be three-yearly inspection (Walshe 2003, p. 63). Evidence of target gaming by the Commission for Health Improvement (2003c and 2004) came also from physical inspections of systems to assure and improve quality of care. Ayres and Braithwaite (1992) observe that it is rare for inspections of nursing homes in the US and Australia to take place without a member of staff giving the inspection team a tip-off of some value. It may be that a visit would have thrown up quality problems such as those in the Bristol heart surgery unit discussed on page 000, lines 00–00, above (where staff were distressed by what was happening), in a way that statistical surveillance on its own could not have done.

However, at the time of writing, if anything, the performance management system has been moving in the direction of widening rather than narrowing the audit hole (Healthcare Commission 2005b). Even though star ratings are due to be abolished, new systems of assessment and inspection emphasize delivery against targets; self-assessment; and surveillance, using readily available data rather than site visits (Healthcare Commission 2005b). These changes, together with the transfer of responsibility for auditing the quality of data in the English NHS from the Audit Commission to the Healthcare Commission (which lacks any physical presence in NHS provider units) suggests less rather than more scope to discover reactive gaming.

None of the measures we propose could be expected to remove gaming completely. But both Soviet history and a broader institutional analysis suggests that they could plausibly be expected to reduce it. And if, as this analysis has shown, there are significant gaming problems in public health care that cannot be prevented by measurement systems that produce a fully robust $M[\alpha_g]$, then

corrective action is needed to reduce the risk of the target regime being so undermined by gaming that it degenerates, as happened in the Soviet Union.

ACKNOWLEDGEMENT

Earlier versions of this paper have been presented at the American Society for Public Administration conference Portland, Oregon, March 2004; the European Conference on Health Economics, London, September 2004; Westminster Economic Forum, London, April 2005. We are grateful for comments from Tim Besley, Carol Propper, David McDaid, Carolyn Heidrich, Jan-Kees Helderman and Rudolf Klein. The usual disclaimer applies.

REFERENCES

Anonymous. 2001. 'Behold, a Shining Light', *Health Service Journal*, 20 December, 14–15.

Anonymous. 2005. 'A&E Survey Highlights Dirt and Waiting Times', *Health Service Journal*, 24 February, 7.

Auditor General for Wales. 2005. *NHS Waiting Times in Wales*. Cardiff: The Stationery Office (http://www.agw.wales.gov.uk/publications/2004/agw2004_9-i.pdf).

Audit Commission. 2003. *Waiting List Accuracy*. London: The Stationery Office (http://www.audit-commission.gov.uk/health/index.asp?catId=english^HEALTH).

Aylin, P. 2003. *Monitoring of Mortality Rates in Primary Care – A Report by Dr Paul Aylin* (http://www.the-shipman-inquiry.org.uk/documentsummary.asp?from=a&id=HP&file=06&page=00001).

Ayres, I. and J. Braithwaite. 1992. *Responsive Regulation*. Cambridge: Cambridge University Press.

Beer, S. 1966. *Decision and Control*. London: Wiley.

Behn, R. 2001. *Rethinking Democratic Accountability*. Washington, DC: Brookings Institution.

Berliner, J.S. 1988. *Soviet Industry from Stalin to Gorbachev*. Aldershot: Edward Elgar.

Bevan, G. and C. Hood. 2004. 'Targets, Inspections and Transparency', *British Medical Journal*, 328, 598.

Bevan, G. and R. Robinson. 2005. 'The Interplay between Economic and Political Logics: Path Dependency in Health Care in England', *Journal of Health Politics, Policy and Law*, 30, 1–2, 53–78.

Bird, S.M., D. Cox, V.T. Farewell, *et al*. 2005. 'Performance Indicators: Good, Bad, and Ugly', *Journal of the Royal Statistical Society*, Series A, 168, 1, 1–27.

British Medical Association. 2005. *BMA Survey of A&E Waiting Times*. London: British Medical Association.

Cabinet Office. 1999. *Modernizing Government* (Cm 4310). London: The Stationery Office (http://www.archive.official-documents.co.uk/document/cm43/4310/4310.htm).

Carlisle, C. 2004. 'How the Government Broke its Bristol Inquiry Pledge', *Health Service Journal*, 4 November, 12–13.

Carter, N., R. Klein and P. Day. 1995. *How Organisations Measure Success. The Use of Performance Indicators in Government*. London: Routledge.

Cole, A. 2001. 'Staying Power', *Health Service Journal*, 3 May.

Commission for Health Improvement. 2001. *Report on the Investigation into Heart and Lung Transplantation at St George's Healthcare NHS Trust*. London: The Stationery Office (http://www.chi.nhs.uk/eng/organisations/london/st_georges/index.shtml).

Commission for Health Improvement. 2002. *Report on the Clinical Governance Review on Surrey and Sussex Healthcare NHS Trust*. London: The Stationery Office (http://www.chi.nhs.uk/eng/organisations/south_east/surrey_sussex/2002/surrey.pdf).

Commission for Health Improvement. 2003a. *NHS Performance Ratings. Acute Trusts, Specialist Trusts, Ambulance Trusts 2002/03*. London: The Stationery Office (http://www.chi.nhs.uk/eng/ratings).

Commission for Health Improvement. 2003b. *NHS Performance Ratings. Primary Care Trusts, Mental Health Trusts, Learning Disability Trusts 2002/03*. London: The Stationery Office (http://www.chi.nhs.uk/eng/ratings).

Commission for Health Improvement. 2003c. *What CHI Has Found In: Ambulance Trusts*. London: The Stationery Office (http://www.healthcarecommission.org.uk/NationalFindings/NationalThemedReports/Ambulance/fs/en).

Commission for Health Improvement. 2004. *What CHI Has Found in: Acute Services*. London: The Stationery Office (http://www.healthcarecommission.org.uk/NationalFindings/NationalThemedReports/AcuteAndSpecialist/fs/en).

Department of Health. 2004. *Chief Executive's Report to the NHS – Statistical Supplement*, May 2004. London: Department of Health (http://www.dh.gov.uk/assetRoot/04/08/26/27/04082627.pdf).

Department of Health. 2005. *Ambulance Services, England*. London: Department of Health (http://www.dh.gov.uk/PublicationsAndStatistics/Statistics/StatisticalWorkAreas/StatisticalHealthCare/StatisticalHealthCareArticle/fs/en?CONTENT_ID=4086490&chk=6NOZfh).

Dunsire, A. 1978. *The Execution Process: Implementation in a Bureaucracy*. Oxford: Martin Robertson.

Ericson, R.E. 1991. 'The Classic Soviet-type Economy: Nature and Implications for Reform', *The Journal of Economic Perspectives*, 5, 4, 11–27.

Farrar, S., F. Harris, T. Scott and L. McKee. 2004. *The Performance Assessment Framework: Experiences and Perceptions of NHS Scotland* (http://www.scotland.gov.uk/library5/health/pafr.pdf).

Goddard, M., R. Mannion and P.C. Smith. 2000. 'The Performance Framework: Taking Account of Economic Behaviour', in P.C. Smith (ed.), *Reforming Markets in Health Care*. Buckingham: Open University Press, pp. 138–61.

Goodhart, C.A.E. 1984. *Monetary Theory and Practice. The UK Experience*. London: Macmillan.

Greer, S.L. 2004. *Four Way Bet: How Devolution Has Led to Four Different Models for the NHS*. London: The Constitution Unit, School of Public Policy, UCL.

Hampton, P. 2004. *Reducing Administrative Burdens: Effective Inspection and Enforcement*. London: HM Treasury.

Heald, D.A. 2003. 'Fiscal Transparency: Concepts, Measurement and UK Practice', *Public Administration*, 81, 4, 723–59.

Healthcare Commission. 2004. *2004 Performance Rating*. London: The Stationery Office (http://ratings2004.healthcarecommission.org.uk/).

Healthcare Commission. 2005a. *Patient Survey Programme 2004/2005. Emergency Department: Key Findings*. London: Healthcare Commission (http://www.healthcarecommission.org.uk/NationalFindings/Surveys/PatientSurveys/fs/en?CONTENT_ID=4011238&chk=0bcNSV).

Healthcare Commission. 2005b. *Assessment for Improvement. The Annual Health Check*. London: Healthcare Commission (http://www.healthcarecommission.org.uk/ContactUs/RespondToAConsultation/CurrentConsultations/fs/en?CONTENT_ID=4016872&chk=61P6R5).

Heinrich, C.J. 2002. 'Outcomes-based Performance Management in the Public Sector: Implications for Government Accountability and Effectiveness', *Public Administration Review*, 62, 6, 712–25.

Holmstrom, B. and P. Milgrom. 1991. 'Multi-task Principal-agent Analyses: Linear Contracts, Asset Ownership and Job Design', *Journal of Law, Economics and Organisation*, **7**, 24–52.

Hood, C. 2002. 'Control, Bargains and Cheating: The Politics of Public-Service Reform', *Journal of Public Administration Research and Theory*, 12, 3, 309–32.

Hoque, K., S. Davis and M. Humphreys. 2004. 'Freedom to Do What You Are Told: Senior Management Team Autonomy in an NHS Acute Trust', *Public Administration*, 82, 2, 355–75.

Howarth, A. 2004. 'Two-hour ambulance delay blamed for teenage boy's death', *The Scotsman*, Monday 18 October.

James, O. 2004. 'The UK Core Executive's Use of Public Service Agreements as a Tool of Governance', *Public Administration*, 82, 2, 397–419.

Kinnel, H.G. 2000. 'Serial Homicide by Doctors: Shipman in Perspective', *British Medical Journal*, 321, 1594–6.

Klein, R.E. 1983. *The Politics of the National Health Service*. London: Longman.

Kornai, J. 1994. *Overcentralisation in Economic Administration*. Oxford: Oxford University Press.

LeGrand, J. 2003 *Motivation, Agency and Public Policy*. Oxford: Oxford University Press.

Litwack, J.M. 1993. 'Coordination, Incentives and the Ratchet Effect', *The Bell Journal of Economics*, 24, 2, 271–85.

Locker, T.E. and S.M. Mason. 2005. 'Analysis of the Distribution of Time that Patients Spend in Emergency Departments', *British Medical Journal*, 10, 1136.

Majone, G. 1996. *Regulating Europe*. London: Routledge.

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

Mayor, S. 2003. 'Hospitals Take Short Term Measures to Meet Targets', *British Medical Journal*, 326, 1054.
Miller, G.J. 1992. *Managerial Dilemmas*. Cambridge: Cambridge University Press.
Moran, M. 1999. *Governing the Health Care State*. Manchester: Manchester University Press.
National Audit Office. 2001. *Inappropriate Adjustments to NHS Waiting Lists*. London: The Stationery Office (HC 452) (http://www.nao.gov.uk/publications/nao_reports/01-02/0102452.pdf).
National Audit Office. 2004. *Improving Emergency Care in England*. London: The Stationery Office (HC 1075) (http://www.nao.org.uk/publications/nao_reports/03-04/03041075.pdf).
Nove, A. 1958. 'The Problem of Success Indicators in Soviet Industry', *Economica* (new series), 25, 97, 1–13.
Nove, A. 1961. *The Soviet Economy*. London: George Allen and Unwin.
Office of National Statistics. 2004. *Regional Trends*, No. 38, Table 7.15 'NHS Hospital Waiting Lists: by Patients' Region of Residence, at 31 March 2003'. London: Office of National Statistics (see also Table 7.15 *Regional Trends*, Nos 35, 36 and 37).
O'Neill, O. 2002. *A Question of Trust*. Cambridge: Cambridge University Press.
Pollitt, C. 1986. 'Beyond the Managerial Model: the Case for Broadening Performance Assessment in Government and the Public Services', *Financial Accountability and Management*, 2, 3, 155–86.
Power, M. 1999. *The Audit Society: Rituals of Verification*. Oxford: Oxford University Press.
Propper, C. and D. Wilson. 2003. 'The Use and Usefulness of Performance Measures in the Public Sector', *Oxford Review of Economic Policy*, 19, 250–67.
Public Administration Select Committee. 2003. *Fifth Report. On Target? Government by Measurement (HC 62-I)*. London: The Stationery Office.
Rowan, K., D. Harrison, A. Brady and N. Black. 2004. 'Hospitals' Star Ratings and Clinical Outcomes: Ecological Study', *British Medical Journal*, 328, 924–5.
Rutstein, D.D., W. Berenberg, T.C. Chalmers, *et al*. 1976. 'Measuring the Quality of Medical Care', *New England Journal of Medicine*, 294, 582–8.
Scottish Executive Health Department. 2003. *Performance Assessment Framework 2003/04* (http://www.show.scot.nhs.uk/sehd/mels/hdl2003_53.pdf).
Secretaries of State for Health, Wales, Northern Ireland and Scotland. 1989. *Working for Patients*. CM 555. London: HMSO.
Secretary of State for Health. 2001a. *NHS Performance Ratings Acute Trusts 2000/01*. London: Department of Health (http://www.dh.gov.uk/PublicationsAndStatistics/Publications/PublicationsPolicyAndGuidance/PublicationsPolicyAndGuidanceArticle/fs/en?CONTENT_ID=4003181&chk=wU4Zop).
Secretary of State for Health. 2001b. *Learning from Bristol – Report of the Public Inquiry into Children's Heart Surgery at the Bristol Royal Infirmary* (the Kennedy Report) (CM 5207(1)). London: The Stationery Office (http://www.bristol-inquiry.org.uk/final_report/).
Secretary of State for Health. 2002a. *NHS Performance Ratings Acute Trusts, Specialist Trusts, Ambulance Trusts, Mental Health Trusts 2001/02*. London: Department of Health (http://www.dh.gov.uk/PublicationsAndStatistics/Publications/PublicationsPolicyAndGuidance/PublicationsPolicyAndGuidanceArticle/fs/en?CONTENT_ID=4002706&chk=dBD1wB).
Secretary of State for Health. 2002b. *The Shipman Inquiry, First Report: Death Disguised* (Chair Dame Janet Smith). London: The Stationery Office (http://www.the-shipman-inquiry.org.uk).
Secretary of State for Health. 2003. *The Shipman Inquiry. Transcript Archive*. Transcript for Day 182 (Tue 14 Oct 2003) (Chair Dame Janet Smith) (http://www.the-shipman-inquiry.org.uk).
Secretary of State for Health. 2004. *The Shipman Inquiry, Fifth Report. Safeguarding Patients: Lessons from the Past – Proposals for the Future*. London: The Stationery Office (Chair Dame Janet Smith) (http://www.the-shipman-inquiry.org.uk).
Shifrin, T. 2001. 'Milburn Puts Managers "on Probation"', *Health Service Journal*, 27 September.
Smith, P. 1995. 'On the Unintended Consequences of Publishing Performance Data in the Public Sector', *International Journal of Public Administration*, 18, 277–310.
Timmins, N. 2005. 'Blair Bemused over GP Waiting Times', *Financial Times*, April 30/ May 1, 2.
Tuohy, C.H. 1999. *Accidental Logics. The Dynamics of Change in the Health Care Arena in the United States, Britain and Canada*. New York: Oxford University Press.
Walshe, K. 2003. *Regulating Health Care*. Maidenhead: Open University Press.

Date received 8 May 2005. Date accepted 6 June 2005.

