

## *Public management by numbers*

# Public service management by numbers: Why does it vary? Where has it come from? What are the gaps and the puzzles?

**Christopher Hood**

*Targets, rankings and intelligence are common forms of public service management by numbers. So what's global and what's distinctively English about this phenomenon? What's new and what's old about the use of targets, rankings and intelligence? And what do we know we don't know about each of these forms of public service management by numbers? The special features of composite ranking systems seem to be a key part of the answer to all these questions*

### **Targets, ranking and intelligence**

Managing government and public services by numbers that describe outputs, outcomes, inputs and throughputs—that is, quantitative performance indicators—is commonly said to be a central theme of contemporary public service reformers. Management by numbers comes in at least three major forms:

- *Target systems*, which measure actual performance against one or more specified aspirational standards expressed as threshold numbers (often, but not always, based on some increment or decrement on what happened in an earlier time-period).
- *Ranking systems*, which measure current or past performance of comparable service units against one another (as information to inform user choice, as information for action by government, or simply as a means for encouraging 'saints' and shaming 'sinners').
- *'Intelligence' systems*, which measure performance for background information (for example as a by-product of administrative processing or complaint-handling), but involve no fixed interpretation of the data in forms such as league tables or comparisons with some stipulated standard.

Each of these basic types can come in different forms, as table 1 overleaf briefly indicates. The three types are often combined into hybrids (for example mixed ranking and target systems,

as we shall see in the three articles that follow). And placing any system in those categories is not always straightforward. For instance, if there are aspirational standards expressed as numbers, but so many such numbers that there is no focus or priority in direction, the boundary between targets and intelligence starts to blur.

### **Putting public management by numbers in its place**

Any serious research programme on modern public services has to go beyond practico-descriptive accounts of such systems to look carefully at the scope and limits of performance metrics in each of those three forms, their intended and unintended effects, and the factors that shape their use. The three articles from the ESRC's Public Services Programme that follow all have light to cast on one or more of those issues, based on analysis of various public service performance indicator systems in Blairite England of the early 2000s.

#### **ESRC Public Services Programme**

The ESRC Public Services Programme is a five-year programme of publicly-funded, peer-reviewed and public-domain research into public service performance. The programme is led by Christopher Hood ([www.christopherhood.net](http://www.christopherhood.net)), it uses methods and disciplinary approaches from across the social sciences, and by the time it ends in 2009 it will have completed over 40 projects and will have held about 50 conferences, workshops and seminars. For more details see [www.publicservices.ac.uk](http://www.publicservices.ac.uk)

*Christopher Hood, Gladstone Professor of Government and Fellow of All Souls College, Oxford, is the Director of the ESRC's Public Services Programme.*

**Table 1. Three applications of performance measurement.**

<i>Application of measures</i>	<i>Basic principle</i>	<i>Simple example</i>	<i>Some variants</i>	<i>Comment</i>
<i>Targets</i>	Stipulated floor standard of performance or change in performance to be achieved within some time period	Percentage efficiency savings or staff reductions required over a budgetary period	Specific targets (applying to individuals or particular organizations) versus global or sector-wide targets	Produce threshold and ratchet effects in behaviour of individuals and organizations subject to targets
<i>Rankings</i>	Data allowing comparison of performance on stipulated indicators among a set of rival units	Sporting leagues	Simple comparisons versus composite leagues (with numbers distilled from other numbers)	Produce output distortions and pressures to change the composition of the league and the nature of the game
<i>Intelligence</i>	Background information	Activity logs, for example of health care 'episodes'	Anonymized data (for example for near-miss reporting) versus attributed performance data	Use is unpredictable by those whose performance is recorded; often combined with targets and rankings

This introductory article aims to step back from that particular time and place, and to put those particular analyses into a broader context by posing three more general questions. First, why does public service management by numbers (particularly for targets and rankings) seem to be practised more in some times and places than in others—and specifically why does a particular form of it seem to have been so heavily emphasized in England in the recent past? Second, what if anything is historically new about the 'public service management by numbers' analysed by the other articles, in the three forms mentioned above? Third, what do we know we still don't know about 'numbers management' in public services?

#### **Management by numbers: a global trend or an English preoccupation?**

The development of quantified performance indicators for managing public services, in one or more of the three forms identified in table 1, is often said to be widespread in the modern world. 'Intelligence' has long existed, in the form of statistics produced to aid policy expertise in domains such as crime, health, demography and public spending. But target systems in various forms are common too. New Zealand became internationally known for developing quantified targets for public service outputs in the late 1980s (as part of a regime that made ministers responsible for 'outcomes' and civil servants for 'outputs'). Since then many other countries have developed more extensive performance indicators in their public services, from the US

Government Performance and Results Act 1993 to France's 'Lolf' budgetary law of 2004. As for rankings, government's statistical 'intelligence' has long provided the basis for comparative league tables, for instance in the various rankings of states and cities in the United States, though such rankings are often compiled by public-interest bodies or firms rather than by government. Official international rankings of public service performance and governance have also developed significantly over the past two decades (see Van de Walle, 2006), though major gaps remain in domains such as transport and crime.

#### **British or English exceptionalism?**

So is there anything distinctive about the UK and specifically England in this supposedly global modern world of quantified performance indicators? Probably not in 'intelligence' applications of performance data, but for target and ranking systems the UK and particularly England seems to have been unusual in at least three ways in recent decades. One is the sheer vigour with which intelligence was turned into central target and ranking systems over that time—Pollitt (2006) found the UK stood out from three other countries in his study in the importance attributed to, and degree of focus on, performance indicators. Indeed, the development of comprehensive PSA targets across British government departments from 1998 arguably took the target approach at the top level of government to a point hardly seen since the demise of the

USSR. A second is the degree to which such uses of performance measures have been dominated by the executive rather than the legislature. For example, the central performance measurement systems in Japan, France and the USA have all been legislatively based, while that applying to England owes little to parliamentary initiative or specific legislation.

Third, the UK and specifically England seem to have been distinctive in the way government-mandated rankings of public service provision have developed within the state. Even in the UK, it is unusual for such rankings to be used to determine funding, though it has been in some cases. For example, university rankings have become commonplace across the world over the past few decades through media or semi-official leagues, but the UK's Research Assessment Exercise (used since 1986 to compare the research performance of all UK universities for the purposes of allocating public research funding) seems to be internationally unusual. It was the first government-sponsored exercise of its kind in the world (following major criticisms of heavily selective cuts made in university funding in 1981 on the basis of undisclosed assessment criteria), and remains the most comprehensive. The UK was also a pioneer in developing from the late 1980s government-mandated leagues of secondary and primary school exam performance (originating in the Education Reform Act 1988, first published for secondary schools in 1992 and for primary schools in 1997), though it was not unique in doing so.

However, official publication of such rankings tables was abandoned in the 2000s in Wales and Northern Ireland (and replaced by information to local parents), half-abandoned in Scotland (in that after 2003 the Scottish executive government provided the raw material for league-tables but did not publish the league itself) and within the UK currently exist in their full-blown form only in England. Moreover, two of the ranking systems—for local authorities and health trusts—that are described in the articles that follow seem to be particular to England and were not emulated even in the other British countries, let alone elsewhere (see Talbot *et al.*, 2005). Many countries have complex formula-funding arrangements for fiscal transfers among levels of government, but England's Comprehensive Performance Assessment regime for local authorities (introduced in 2001 and determining how heavily they are regulated

from the centre) is a distinctive approach to composite ranking of local authority performance. The same goes for the officially-mandated star rating system for National Health Service (NHS) hospitals in England that ran from 2001 to 2005.

### **Explaining English exceptionalism?**

So if there *is* something special about 'public management by numbers' in England, it is those three features that seem to encapsulate it. For some, such developments amount to 'best practice', an 'English cure' for everyone's public service problems, while others see them as an English disease that others should strenuously try to avoid. But leaving aside the cure-or-disease issue (dealt with by two of the articles that follow), how can we explain the relatively heavy emphasis in England on top-down target systems in the recent past and particularly on the centrally-mandated league-table form of public service management by numbers?

### **Scale and centralization**

The most likely explanation (following a line of analysis also offered by Pollitt, 2006) would seem to be some mixture of three closely interlinked features, namely scale and centralization, institutions and culture. As for the first, England has often been said to be the most centralized country in Europe in the sense of having no elected levels of government between the UK 'Imperial' parliament and local authorities. Indeed, in health care, while other countries have followed the 'Beveridge model' in providing universal coverage, sub-national government plays an important role in the Scandinavian countries and Italy, in contrast to England where the Department of Health can reorganize the NHS at will. Scale creates conditions in which central government has both the motive and opportunity to develop elaborate target and league table systems in a way less likely to apply in less centralized countries. It creates the conditions for a more developed managerial transfer market for players such as health trust chief executives or school 'superheads' than can exist in smaller leagues. And it creates a degree of 'relational distance' between those heading delivery agencies and the central establishment that allows for the use of 'terror' in measured performance systems in a way that is harder in smaller societies with more tightly-linked and overlapping political and social élites (see Hood and Bevan, 2006). Such factors may help to account for the fact that within the UK it is only

in England that health care performance measures were used in the 2000s both as targets (to reward managers who performed well and as a trigger for sacking under-performing managers and sending in 'turnaround teams') and as composite ranking systems that helped to determine trusts' 'foundation' status, involving the ability to borrow money on their own account.

### **Institutions and culture**

A similar motive-and-opportunity point could be made for two other institutional features often said to be distinctive to the UK as a whole. These are the strength of the central coordinating departments within central government and the peculiarity, noted above, of the UK's NHS in an international context, as a public health care system traditionally operating as a centrally organized public-bureaucracy system of providing health care, in contrast to the insurance-based and variously-provided systems of health care in much of the rest of the world. Performance data of ever-increasing elaboration came to be collected after the creation of the NHS in 1948 (Jowett and Rothwell, 1988, pp. 5–6), but arguably the opportunity to use those data for heavy-duty ranking exercises, first floated in the early 1980s, came only at a moment of *perestroika* from the 1990s, with the weakening of a long-standing implicit bargain between doctors and the government, in which 'while central government controlled the budget, doctors controlled what happened within that budget' (Klein, 2001, p. 64).

### **Management by numbers: a new era—or history denial?**

Public service management by numbers is often presented as a new phenomenon, and part of a relatively new way of thinking about government and public services. Indeed, the impression of novelty, combined with the often-noted rhetorical power of numbers (Maguire, 1994, p. 236), seems to be a key part of the appeal of an approach to public management heavily focused on performance indicators. But some of that novelty can be exaggerated. Indeed, each of the three uses of quantitative performance measurement in public services considered here has a history both in theory and in practice (see also Jowett and Rothwell, 1988).

### **Targets**

The use of performance indicators as target systems no doubt has an earlier history, but is

conventionally associated with Frederick Winslow Taylor's approach to 'scientific management' by setting production quotas linked to individualized payment systems—an idea first developed in the 1890s and turned into a prescription for 'government efficiency' in 1911 (by measurement of output linked to production targets), which was published shortly after Taylor's death (Taylor, 1916). It became central to Soviet management and economics after Lenin's famous embrace of Taylorism as a management system in an article published in the Bolshevik newspaper *Izvestiya* in 1918 (Lenin had denounced Taylor's approach as a form of human 'enslavement' before the 1917 revolution) (Merkle, 1980). And Soviet experience with the target approach to economic management after 1928 led in time to relatively sophisticated discussions among economists about the design of target systems and consequences such as threshold effects, ratchet effects and output distortions. Quantitative target systems were also used in Britain for the management of munitions and other war production in the 20th-century world wars, and have been used to manage case-handling in welfare and job-placement bureaucracies for half a century at least, with classic studies of the process dating back to the 1950s (see Blau 1955; Jowett and Rothwell, 1988).

### **Rankings**

The use of performance indicators as rankings also has a long history. The prescriptive idea can be traced back at least as far as the philosopher and social reformer Jeremy Bentham's late 18th-century call for what would now be called performance accounting in government bodies, linked with what he called the 'tabular-comparison principle' (league tables, in modern parlance: see Hume, 1981, p. 161). International rankings of public services, for example on naval strength and crime rates, can be traced back a long time too. International crime statistics are conventionally dated from the General Statistical Congress held in Brussels in 1853. In the case of naval strength, after warships became a specialized kind of ship (a process complete in Europe by the 17th century), they could be counted to produce indicators of the naval capacity of the principal powers, and that had become formalized by the early 20th century. (The annual publication of comparative naval strength, *Jane's Fighting Ships*, originally compiled by John Frederick Thomas Jane, dates from 1898.)

## Intelligence

The use of quantitative performance indicators as 'intelligence', background information collected for managers or policy-makers to review but not necessarily linked to target or league table systems, probably has an even longer history. After all, measurement of forest production to manage state forests at the maximum sustainable yield goes back to 18th-century scientific forestry (Scott, 1998). Crime statistics were published in Britain from the mid-19th century, crime clear-up rates have been used as an indicator of police performance in the US and other countries for many decades, and proportionate collection costs have been used as a performance indicator in tax administration for longer than that. The famous British nurse, hospital administrator and statistician Florence Nightingale developed a system of detailed statistics about hospital performance in the 1840s when she served at Scutari during the Crimean War and developed graphical methods of quality control said to be far more extensive than those in use a century later (Cohen, 1984; Wadsworth *et al.*, 1986; Jowett and Rothwell, 1988, p. 5). However, service-wide performance indicators were collected in the British NHS from 1948 and Jowett and Rothwell (1988, pp. 10–12) list 68 such indicators applying to district health authorities in 1983.

## Are we in history denial?

Such cases suggest that at least some aspects of public service management by numbers are 'modern' only in the sense of the European scholarly convention that sees modern history starting several centuries ago. So does the idea that there is something new about public management by numbers represent 'history denial,' as so often happens in government and public services? If so, what would account for such denial? And, if not, what precisely is different about quantitative performance measurement and management of public services in the modern age?

If some sort of history denial was going on about the earlier life and times of public service management by numbers, the innocent explanation would be that of simple ignorance of earlier phases in the management by numbers movement on the part of those leading the movement today.

A less innocent explanation might be that such history is ignored less out of simple ignorance than because it is inconvenient and ill-suited to the rhetorical purposes of today's reformers. After all, the fact that earlier forms of management by numbers were abandoned, or proved

problematic, inevitably prompts questions as to what shortcomings might have led to that result, and can undermine an unstoppable 'wave of history' view of such developments. Much of the historical experience with the use of target systems comes from the Soviet Union, which was arguably the wrong kind of historical precedent for public service reformers to invoke in turn of the 21st-century capitalist democracies (and for that very reason has tended to be invoked by those critical of the management by numbers movement). Even the non-Soviet history, such as the experience with using target systems for aircraft production in the Second World War, tends to be 'the wrong kind of history' in that it shows up the limitations and inherent dilemmas of management by numbers that latter-day advocates might be reluctant to highlight.

There might well be something in both explanations. But a more defensible explanation might be that such history is not relevant because of the qualitative difference between today's public management by numbers and that of yesteryear. Such an argument cannot plausibly be applied to the use of performance measures as 'intelligence,' and even for target applications the plausibility of such arguments is debatable. After all, as shown above, target systems of one kind or another have been around in public management for a long time. And even in 'targetworld' Britain of the 2000s, the much-debated PSA target system for health developed under Tony Blair's government after 1998 had far fewer components at the topmost level than the earlier Conservative model that it replaced. What seems to be different about the recent past is the greater top-level political salience given to measured performance targets, their use in modern capitalist democracies rather than war economies or Soviet-type systems, their specific application to public services rather than the economy-wide targets used in the post-Second World War era of indicative planning (notably by the French Commissariat General du Plan or the short-lived 1965 British National Plan that imitated it) and their extension from lower-level executive bodies and the middle level of management in case-handling bureaucracies to policy departments. Some of those changes—and the associated organizational routines and institutional developments—certainly involved innovation. But, significant as such developments may have been, they were further steps on a path that had been heavily trodden before.

## Composite rankings: a new development?

However, for rankings, and particularly for the composite rankings that were argued earlier

to be a distinctive feature of England's approach to public management by numbers in the early 2000s, the qualitative difference argument seems more plausible. As shown above, there is an earlier history of the use of performance indicators in rankings, and various international leagues and semi-official rankings of cities or states. But complex league tables of the kind discussed in two of the articles that follow (that is, systems that create rankable numbers by mixing together a host of individual indicators with different weightings into a single score) do seem to represent a rather new development. That applies both at the international level of rankings, with complex composite numbers such as the competitiveness index and the World Bank's governance ratings, and at the national level with composite indicators of organizational performance such as those discussed in the previous section. No doubt such developments are partly a product of technological change and have been facilitated, even maybe enabled, by the computer age, though as argued in the previous section their particular application in the England of the 2000s seems best explained by a mix of scale, culture and institutions. The complexity and opacity of such systems (albeit ironically often claimed to bring greater transparency into funding and performance assessment) seems to take measurement into rather new territory, and it is significant that two of the articles that follow are concerned with composite performance indicator systems.

#### **Management by numbers: what we know we don't know**

In one sense, the phenomenon of public management by numbers in the three ways discussed earlier hardly seems to be an understudied phenomenon. After all, practitioner and public management journals like this one have long been awash with accounts of the latest phases in the development of such systems, and general debates about their efficacy or otherwise as a way of steering complex delivery systems. But in spite of that, there are at least three important things that we know we don't know about performance measures as targets, rankings or intelligence.

#### **How valid and reliable are complex composite measures?**

First, while much has been written on performance measurement in general, we know relatively little about the validity and reliability of complex composite performance measurement systems. Two of the articles that follow report on

the hard task of assessing the validity and reliability of composite indicators of public service performance—work that involves complex and technical analysis of large datasets, and as Rowena Jacobs and Maria Goddard show, in some cases requires simplified forms of those datasets to be laboriously assembled before they can be tractable to analysis even in an age of super-computers.

Measurement error is unavoidable in any attempt to quantify. It arises from several sources, including:

- Simple mistakes (clerical error, such as inadvertent double-counting or omissions at the source of data collection).
- Sampling error (the indicator, time-period or subunit taken is not representative of the overall population).
- Categorization errors (where perplexity about how to fit cases into categories may result in faulty assignment of those cases).
- Gaming or cheating (deliberate massaging or outright fabrication of numbers collected with the intention of improving the position of an individual or organization).

While the simple mistake form of measurement error arises in all of the three uses of performance measures considered here, and sampling error and categorization issues may be equally common to them all too, the gaming form of measurement error can be expected to be highest for targets and rankings, especially of the published type, and correspondingly less for measures in the form of 'intelligence'. But we know relatively little about the extent of gaming or cheating in target or ranking systems, or indeed about where the culture draws the lines in practice between what is seen as gaming and what as cheating (a question that needs an ethnographic approach to answer directly—see Hood, 2006).

#### **Targets, rankings or intelligence?**

Second, we do not have coherent theories, whether normative or descriptive, official or academic, about what social conditions match up with what kinds of performance indicators in public service management. If such indicators can be used as targets, as rankings or as 'intelligence', what conditions are appropriate to each of these applications? That question can be posed in instrumental or managerial terms, or in sociological or historical terms, to answer the questions raised earlier about what might be distinctive about England in the early 2000s, or to explore whether the development of performance indicators in the three forms considered here describes some

**Table 2. Targets, rankings or intelligence: when to use what?**

<i>Intended effect</i>	<i>Use performance indicators as:</i>		<i>Limits</i>
Raising a limited number of standards	Targets		Ratchet and threshold effects; gaming added to other sources of measurement error
Sweating and stretching		Rankings Activity logs	Statistical noise; output distortion; gaming added to other sources of measurement error
Developing learning capacity and diagnostic power—adding knowledge for uses that may not be fully foreseen			Intelligence
			Lack of transparency and clear incentives

sort of linear historical development or something more circuitous and cyclical.

One possible starting-point for a contingent or instrumental theory of purposes that might fit the three types of performance indicator considered here would be to distinguish between raising basic levels of performance, sweating and stretching public service provision systems, and serendipity, or building a knowledge base about public service provision to be used in ways that may not be predictable. These three objectives are summarized in table 2. If the intention is to put the focus on baseline standards below which performance (or performance improvement) should not fall, for example in speed or accuracy of treatment, targets are the most direct way of achieving that policy goal. For example, it seems inconceivable that the massive reduction in waiting times for hospital treatment in England since 2001 (with targets for first outpatient appointment and elective inpatient admission set at six and 18 months for 2001, and down to an 18-week target for admissions following GP referral by 2008) could have been achieved by publishing rankings or collecting waiting time data simply as intelligence. But in a no-free-lunch world, target systems have well-known costs as well, typically in the form of ratchet effects or threshold effects (or conceivably both, though commonly the more we try to avoid ratchet effects, the more we will create threshold effects and vice versa), and these effects may well become serious as time goes on.

On the other hand, if the intention is to 'sweat' assets or put broadly comparable service providers under pressure to do as much as they can without specifying floors or ceilings (and thus conveniently avoiding ratchet and threshold effects), then rankings are the obvious application of performance measures.

But ranking systems come at a cost as well. They are known to be vulnerable to statistical noise, as Jacobs and Goddard show (see also Goldstein and Spiegelhalter, 1996; Marshall and Spiegelhalter, 1998). Like target systems, they are likely to produce output distortions as producers learn to find ways that move their organizations up the league-tables in ways that do not reflect the intentions of those who framed the rankings, or ignore non-measured activities. Both processes are at the heart of criticisms of the way schools play league-table games (for instance, by focusing on subjects originally intended for adult learners that counted as several good GCSE grades) and universities play league-table games in the RAE (for instance by ignoring outputs not rated but nevertheless important in the scholarly profession, such as book reviews).

If the intention is to improve background knowledge or develop expertise about the working of a system without creating strong pressures for gaming that distort the reported numbers, 'intelligence' will be the tool of choice. Indeed, an intelligence approach may be what performance indicator systems rebound to after gaming pressures have distorted target or ranking systems. Moreover, agencies that run target or ranking systems will often need to have 'intelligence' performance indicators as well, given the pressures for distortion on the former type (for instance, the World Bank practises such an approach). If gaming is an issue, intelligence has the advantage of unpredictability: since managers do not know what indicators will be used with what weighting for what purposes, their incentive and ability to game the numbers will be correspondingly reduced. But intelligence also has the disadvantages that go with unpredictability, including multiple possible interpretations, and lack of transparency and

clear incentives for public service providers to pursue consistent and clearly stated goals.

### Unintended effects

Third, while there is a rich literature on the unintended effects of target systems (mainly in terms of the unintended distortion of managerial effort they can produce), we know rather less about unintended effects of league-table or intelligence systems, or even about the broader unintended effects of target systems. Such unintended consequences can be looked at through many different analytic lenses, and the recent history of public management by numbers in the UK and particularly England is rich in swiftly-acting examples of such effects. Cases include the contribution of the 2001–2005 hospital star rating system in England to the major NHS deficit which came to light early in 2006 (hospital trusts could achieve a one-star rating without hitting financial targets—see Bevan, 2006) and the contribution of the target system to the 2006 political crisis in the Home Office over release of foreign prisoners into the community without deportation (by focusing the energies of senior administrators into targetized activities and ‘low hanging fruit’ such as the easy deportation cases). Even cases of this kind show that measured performance systems can produce unintended consequences that involve not just passing embarrassment but serious casualties at the top of government.

However, beyond such fairly quick-acting and politically concentrated unintended effects of measured performance systems are unintended consequences that are long term and system-level, and almost by definition we know very little about such effects. But that does not mean they are not there. Will a sustained emphasis on public service targets with harsh sanctions for failure lead to the sort of system collapse some scholars associate with the cumulative effect of the USSR’s target system (Braguinsky and Yavlinski, 2000)? Will heavy emphasis on public service performance numbers expressed as high-stakes targets and rankings lead to a further loss of public trust in government statistics, through the perception that for every set of statistics showing good (or bad) performance there is an equal and opposite set of statistics pointing in the opposite direction? Will instruments like university research rankings unintentionally lead to a long-term decline in research quality through pressures to play safe or produce short-term publications to fit with administrative census dates? Such questions go rather beyond the

scope of the three articles that follow. But that does not mean they are not important. ■

### References

- Bevan, R. G. (2006), Setting targets for health care performance: lessons from a case study of the English NHS. *National Institute Economic Review*, 197, pp. 67–79.
- Bevan, R. G. and Hood, C. (2006), What’s measured is what matters: targets and gaming in healthcare in the English public health care system. *Public Administration*, 84, 3, pp. 517–538.
- Blau, P. M. (1955), *The Dynamics of Bureaucracy* (Chicago University Press, Chicago).
- Braguinsky, S. and Yavlinski, G. (2000), *Incentives and Institutions: The Transition to a Market Economy in Russia* (Princeton University Press, Princeton).
- Cohen, I. B. (1984), Florence Nightingale. *Scientific American*, 250 (March 1984), pp. 128–137.
- Goldstein, H. and Spiegelhalter, D. J. (1996), League tables and their limitations. *Journal of the Royal Statistical Society, Series A*, 159, pp. 385–443.
- Hood, C. (2006), Gaming in targetworld. *Public Administration Review*, 66, 4, pp. 515–522.
- Hume, L. (1981), *Bentham and Bureaucracy* (Cambridge University Press, Cambridge).
- Jowett, P. and Rothwell, M. (1988), *Performance Indicators in the Public Sector* (Macmillan, London).
- Klein, R. (2001, orig. 1983), *The Politics of the National Health Service*, 4th edn (Prentice-Hall, Harlow).
- Maguire, M. (1994), Crime statistics, patterns and trends. In Maguire, M. *et al.* (Eds), *The Oxford Handbook of Criminology* (Oxford University Press, Oxford).
- Marshall, E. C. and Spiegelhalter, D. J. (1998), Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *British Medical Journal*, 316 (7146), pp. 1701–1704.
- Pollitt, C. (2006), Performance management in practice: a comparative study of executive agencies. *Journal of Public Administration Research and Theory*, 16, 1, pp. 25–44.
- Scott, J. (1998), *Seeing Like a State* (Yale University Press, New Haven).
- Talbot, C. *et al.* (2005), *Exploring Performance Regimes: Comparing Wales with Westminster, a Report for the Wales Audit Office* (CPPM, Manchester Business School).
- Taylor, F. W. (1916), Government efficiency. *Bulletin of the Taylor Society* (December), pp. 7–13.
- Van de Walle, S. (2006), The state of the world’s bureaucracies. *Journal of Comparative Policy Analysis*, 8, 4, pp. 437–448.
- Wadsworth, H., Stephens, K. and Godfrey, A. B. (1986), *Modern Methods for Quality Control and Improvement* (Wiley, New York).