

---

## *Public Management by Numbers*

---

Editorial: Public Management by Numbers <i>Andrew Gray and Christopher Hood</i>	92
Debate: The Civil Service—The Way Forward <i>Sir Gus O'Donnell</i>	93
Public Service Management by Numbers: Why Does it Vary? Where Has it Come From? What Are the Gaps and the Puzzles? <i>Christopher Hood</i>	95
How Do Performance Indicators Add Up? An Examination of Composite Indicators in Public Services <i>Rowena Jacobs and Maria Goddard</i>	103
The Perils and Pitfalls of Performance Measurement: The CPA Regime for Local Authorities in England <i>Iain McLean, Dirk Haubrich and Roxana Gutiérrez-Romero</i>	111
PMM Instructions for Authors	118
Performance, Strategy and Accounting in Local Government and Higher Education <i>Martin Broad, Andrew Goddard and Larissa Von Alberti</i>	119

**Reprint of *Public Money & Management's* ESRC theme edition articles guest edited by Christopher Hood. For more information about PMM see <http://www.cipfa.org.uk/pt/pmm>. Or contact the Managing editor Michaela Lavender: [michaela.lavender@cipfa.org.uk](mailto:michaela.lavender@cipfa.org.uk)**

# Editorial: Public Management by Numbers

**Andrew Gray and  
Christopher Hood**

One of the more ugly expressions used by managers in the National Health Service these days is 'hanging offences'. These are the career-threatening failures to meet the most significant performance targets set by government. You do not have to look far to gauge their significance: the bodies of the convicted litter the field of primary care in the aftermath of last year's reorganization.

Other services have their equivalents. They have quantitative expression in common. But they are just the visible tip of the formidable iceberg of quantitative performance measurement that has grown over the past quarter of a century and which is now a dominant feature of the seascape of public services management in this and many other countries.

*Public Money & Management* has charted many of these developments in both its thematic editions and individual articles. Indeed, the subject has been one of the most common in submitted articles. The interest reflects its importance: the huge amounts of public service activity and expenditure require for effective governance a valid, reliable and timely method of measurement. Yet the contributions have often questioned what to some has seemed a remorseless growth of an unevaluated industry:

- What is the assurance that the numbers validly represent the end-users' experience of the

- performance?
- When are different formats such as numerical performance data and league tables appropriate?
- How can we protect against developing an emphasis on provider and managerial gaming?
- Why do some organizations emphasise performance data and their management by building their organizational processes around them, while others marginalize the whole business?

The Economic and Social Sciences Research Council has recognized these and related questions as central to an assessment of performance metrics in contemporary public services. They are an important component of its current research programme on public services. From this programme we publish a selection of articles looking at the metrics issue from different, but complementary, angles. We begin with an analysis by Christopher Hood of the performance by numbers business including its targets, rankings and intelligence systems. This article is followed by contributions by Rowena Jacobs and Maria Goddard on the use of composite indicators; by Iain McLean, Dirk Haubrich and Roxana Gutiérrez-Romero on local authority comprehensive performance assessment; and by Martin Broad, Andrew Goddard and Larissa Von Alberti on the organizational management of performance measurement.

Together the articles develop our understanding of and raise questions about the technical, organizational and political dimensions of performance measurement. However, there is more than a hint of paradox: that, in theory and practice, we are coming to regard performance measurement as an obstacle as much as a facilitator of performance delivery. ■

# The Civil Service— The Way Forward

**Sir Gus O'Donnell**

The role of the state has been redefined many times over the past 100 years, responding to changing social and economic pressures. The civil service has constantly evolved and adapted to deal with this, from the number of people it employs, to the very shape and structure of Whitehall.

We are now at a pivotal point in the civil service's history where we face a number of unknowns. Many key challenges—security, migration, the environment—are global in nature. Moreover, citizens' expectations of the state have risen and the very relationship between citizens and the government has changed and continues to do so. As the government responds to these challenges, so must the civil service. And our response needs to go further than transformation in numbers and structures.

The interaction between citizens and government is a useful starting point for understanding the terms of this debate. Within a generation, the process by which citizens tell government what it wants, through to change or action on the ground—to borrow a concept from economics, the 'transmission mechanism'—has changed dramatically. Traditionally this has been through political mechanisms: local and national elections, membership of political parties. Given declining voter turnout and decreasing party membership, it has been said that citizens today are not as interested in issues that affect society as previous generations. The truth is perhaps more complex. What is evident is rising interest in single issues (the mobilization of people around the 'Make Poverty History' campaign is a good example), coupled with confidence in alternatives to conventional transmission mechanisms to affect change.

The diversity of communication channels has implications for the state's ability to reach its citizens as well as the other way round. Whereas once it was easy to communicate with a mass audience through a limited number of television stations and newspapers, the landscape today is far more dispersed, intangible and immediate. This means that the state will have to be much more creative in how it connects with citizens

to understand their needs, something that is already starting to happen.

The internet has also contributed to a shift in power between citizens and the state that poses further challenges for the government and civil service. Easy access to information on public services, for example school league tables and hospital standards, has made for a more assertive, confident citizen who is prepared to question specific aspects of the government's performance. This status of citizens as consumers, rather than receivers, of public services is a relatively new development which the present government has advanced through the choice and personalization agenda. In addition, there is also evidence showing that citizens' expectations of the role of the state are increasing. For example, more people today think that the state should be responsible for keeping prices under control and provide the means by which people can get jobs, than 10 years ago.

However, while users' demands of public services are higher than ever, the job of government is arguably getting harder. This is particularly the case when compared to the job of private sector counterparts against which the state's performance is increasingly judged. For a start, government has a duty to deliver for all members of society—however vulnerable and hard to reach—not just for those who can pay. But changing demographics coupled with fiscal constraints means that in the years ahead the state will have to provide for more people, at higher cost per head, but with fewer resources.

How does the state meet all these challenges? A simple solution might be to increase the resources at its disposal, for example through taxation. Political issues aside, it is not clear that in today's globalized economy this is a realistic option. The increasing influence of the emerging market economies, in particular China (whose economy overtook the UK's last year), not only has an effect on public demand for better products and services at cheaper prices, but also has implications for government's fiscal and wider economic stance.

The civil service is acutely aware of the need to do more with less, which is why efficiency is so high on our agenda. Even during the past few years of fairly rapid growth in public spending, the civil service still grew more slowly than overall employment. There are 14,000 fewer civil servants than a year ago and numbers will

continue to fall.

Securing greater efficiency is just one part of a wider programme of civil service reform which aims to improve the way we work, where we work and how we engage with our delivery partners and the wider public. One element of this reform programme that will be a major force for change is Capability Reviews. These provide a systematic way of identifying departments' development needs in terms of strategy, leadership and delivery. The reports setting out the findings from the first and second tranches of departments reviewed have been published and action is already being taken forward as a result of this.

As well as priorities for individual departments, Capability Reviews are also identifying challenges that cut across the civil service as a whole. We are already aware of some of these issues and have steps in place to address them. For example, the Professional Skills for Government programme is ensuring that our policy-makers have the right mix of skills and experience to design appropriate and informed policy solutions.

Looking forwards, the drive to professionalize our workforce needs to continue and specifically focus on 'commissioning' skills. This is necessary because there are a growing number of areas—prisons, infrastructure, education—where the state is drawing on private and third sector providers. The civil service needs to ensure that it has the right people with the right skills in place to manage these relationships. In addition, we will need to

improve our understanding of the various models for delivering public services at a strategic level. For example, the Department for Education and Skills delivers virtually nothing directly, but operates through wider public sector, voluntary and private partners, whereas the Department for Work and Pensions administers all of its services through the civil service. Civil servants need take informed decisions over what delivery model works best in what circumstances.

So much for the immediate challenges, what of the role of the state in the longer term? How will and should the civil service adapt? We are certainly going to see a different, strategic, state in the future, where there will be a strong partnership between the citizen and the state and where power will be devolved. It will be a state that will shape rather than control the behaviour of its citizens to achieve its aims, for example on the public health and obesity agenda. Alongside this, we can expect to see a different civil service. This new civil service will remain true to its underlying values of honesty, objectivity, integrity and impartiality, but these values will be fused with the '4 Ps': pride, passion, pace and professionalism. Together these can keep the civil service relevant and effective and bring the dynamism needed to deliver 21st-century public services. ■

*Sir Gus O'Donnell is Cabinet Secretary and Head of the Home Civil Service. This article is based on a lecture that Sir Gus gave to the Public Management and Policy Association in February 2007.*

# 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.

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

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

### **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)

**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

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 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

**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

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).

# How Do Performance Indicators Add Up? An Examination of Composite Indicators in Public Services

**Rowena Jacobs and Maria Goddard**

*Composite indicators are an aggregation of underlying performance indicators into a single index and have been used widely in the public sector to create league tables. This article investigates the degree to which composite measures are an appropriate metric for measuring performance. The authors illustrate the degree of uncertainty in the construction of composites and how rankings are sensitive to the way in which the performance indicators are aggregated. The article highlights the issues which need to be considered in the development and use of composite indicators for performance management purposes.*

In almost every newspaper or government report we read there is some reference to the performance of public sector organizations. The public have become used to judging schools, universities, hospitals, local councils, and other public sector organizations, in terms of their performance rating. This has led to a proliferation of league tables. Managers have become used to being held accountable for the performance of their organization and achieving good ratings is a major endeavour. The use of league tables and rankings are common practice in 'management by numbers' (Hood, 2007).

Although a variety of performance measures exist, current government policy in England emphasises the creation of composite indicators (adding up a number of indicators to form an aggregate indicator) and they are used widely in health, social services, education, universities, local government and other service areas (Freudenberg, 2003; Joint Research Centre, 2002). Whether it is the 'star ratings' of hospitals or social service departments or the 'research assessment exercise' ratings of universities, the use of a single score or label that summarises a wealth of underlying performance data, has wide appeal. These composite performance ratings have taken on great importance as they are often used to reward or penalize organizations.

Little is known about the degree to which composite measures are an appropriate metric for evaluating performance in the public sector.

Do they reflect accurately the performance of the organization? To what degree are they influenced by the uncertainty surrounding the underlying indicators on which they are based? Are they robust or are they subject to instability depending on the way in which they are constructed? Many of these issues are of course common in considering any type of performance indicator. However, the use of composite measures compounds these difficulties and is associated with additional methodological challenges that influence the degree to which they may represent an adequate performance measure.

Why is there such a compelling need to produce a summary indication of performance in the form of a composite score? Possible arguments in favour of creating composites include focusing attention on important policy issues, offering a more rounded assessment of performance and presenting the 'big picture' in a way which the public can understand. In contrast to piecemeal indicators based on individual performance measures, they can offer policy-makers a summary of complex multi-dimensional issues. They can of course still be supplemented by more detailed performance information and can be used to define standards and what constitutes 'success' and 'failure'. They provide an attractive option for accountability purposes, as it is easier to track progress of a single indicator over time rather than a whole package of indicators. Although there is a range of counter-arguments

*Rowena Jacobs is a Research Fellow at the Centre for Health Economics, University of York.*

*Maria Goddard is Professor of Health Economics and Deputy Director of the Centre for Health Economics, University of York.*

(Smith, 2002), the temptation to summarise complex processes into a single figure seems irresistible.

In this article we examine whether composite indicators are a good way of measuring performance in the public sector. In particular, we look at whether such measures are robust and can accurately reflect genuine differences in performance. We use a generic composite measure to illustrate that the methods adopted and the judgements made in constructing the composite can have a profound impact on the ranking of organizations, and hence on the incentives faced by public sector organizations.

## Data and Methods

### Data

We used data on composite performance measures from two key public services in England: healthcare and local government (Healthcare Commission, 2005; Audit Commission, 2005). The data comprise:

- NHS trust star ratings for around 180 acute NHS trusts from 2000/01 to 2004/05 covering around 40 performance indicators.
- Comprehensive Performance Assessment (CPA) ratings for around 150 local authorities from 2001/02 to 2004/05 covering around 110 best value performance indicators (BVPIs).

### Methods—Constructing a Generic Composite

Our methodology was to construct a generic composite indicator for each of the two sectors, using them to examine the sensitivity of rankings of organizations to the various methodological issues involved at each step in the construction. The purpose was not to replicate the ratings that currently exist in these sectors but, rather, to explore the sensitivity of a generic composite to the various methodological judgements involved in its construction.

In order to make this generic measure practical, but also realistic, we selected the underlying indicators for each composite applying the following criteria:

- The indicators should cover the broad range of performance (we used factor analysis to obtain the key dimensions).
- The indicators should not have large numbers of missing values.
- They should exhibit stable statistical properties.
- The indicators should preferably be available for more than one year.

This resulted in a subset of 10 performance indicators for healthcare and a subset of 35 performance indicators for local government. All indicators were transformed to 'more is better', for example death rates were converted to survival rates. We then standardized the performance indicators so that they could be aggregated.

In the first instance, we simply added the indicators up for each hospital and local authority applying an equal weight to each indicator to create the composite index and a ranking based on the index. Hospitals and local authorities with missing data were excluded, giving a final sample of 117 hospitals and 97 local authorities.

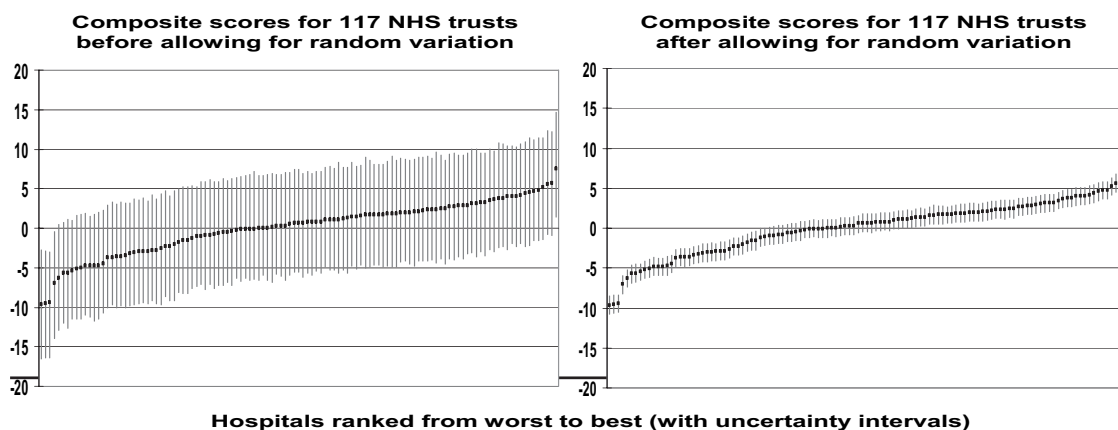
### Examining Uncertainty

There is uncertainty in all data and all performance measures will be subject to random fluctuation and measurement error. We tried to address this by estimating the magnitude of such random fluctuation and simulating the consequent impact on performance ratings. Without the use of simulation methods we would only be able to calculate a single composite index for each hospital or local authority (Mooney, 1997). With simulation methods, we can explore the impact of uncertainty on the composite measure. We performed 1,000 simulations for each hospital and local authority by drawing random samples from performance indicators with the same statistical properties as the original sample. This enabled us to calculate 1,000 composite indicators and rankings for each hospital and local authority to examine the range of performance scores a single organization obtains.

### Examining Random Variation

Every performance signal will consist of real variations in performance, variations in performance which may be the result of differing local circumstances, and random fluctuations (natural statistical variation). It is important to be able to disentangle different sources of variation on performance measures because we are interested in isolating the genuine performance variation that is within managerial control. For each of the performance indicators, we sought to estimate the proportion of variation caused by factors such as measurement error and random variation. We used the longitudinal data on each underlying performance indicator and used regression methods to isolate the variation *within* organizations over time (Wooldridge, 2002). This was used as an estimate of the

Figure 1. Composite score and 95% confidence interval for 117 NHS trusts.



proportion of that indicator's variation that is random and thus beyond the control of the organization. The remainder of the variation was assumed to reflect genuine variations in levels of performance.

#### *Examining Alternative Aggregation Methods*

In practice, a range of different aggregation methods can be used to construct composites. Different weighting structures can be applied to reflect different priorities attached to attainment on certain indicators. There is little consensus though on whose preferences these weights should reflect, or how these preferences should be elicited (Mullen and Spurgeon, 2000; Dolan *et al.*, 1996). Decision rules or algorithms are also often applied as an aggregation method to ensure attainment of minimum standards on some indicators, thus introducing implicit weights. Decision rules may, for example, assign  $x$  score if the organization is rated good on three out of four indicators; or  $y$  if rated good on only two out of four indicators.

We tested a range of alternative approaches to aggregating the individual performance indicators into a composite:

- Adding the 10 indicators (healthcare) or 35 indicators (local government) in a linear fashion with equal weights (base case).
- Adding the indicators but varying the weights.
- Applying non-linear decision rules to assign hospitals or local authorities to ordinal categories (for example 0–3 stars or 'excellent', 'good', 'poor'). These types of sequential rules are commonly applied in aggregation methods (for example in the star ratings and CPA).

While actual composites rarely take account of

uncertainty or random variation, they do often apply different aggregation methods or decision rules. The methods applied here and the examples tested are therefore as realistic as possible, within the constraints of our criteria for selecting the indicators and creating the original composite.

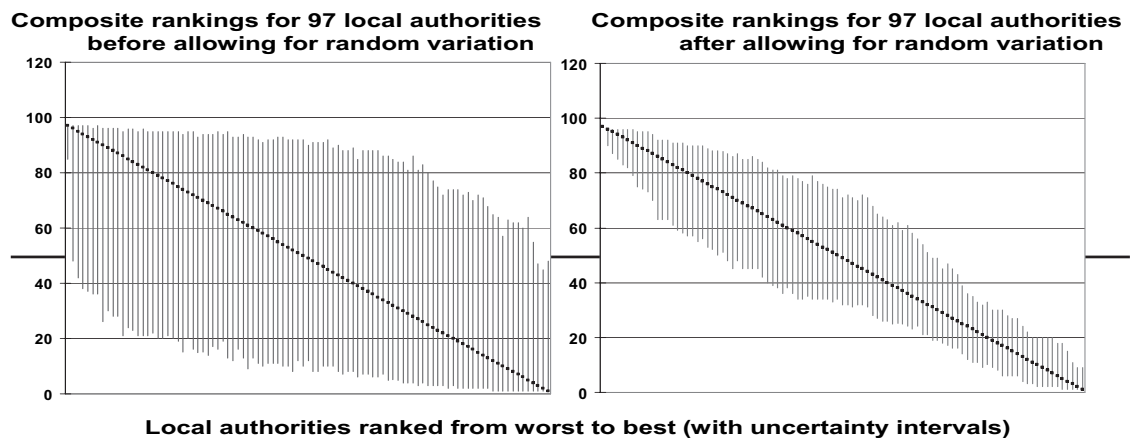
## **Results**

### *Examining Uncertainty and Random Variation*

Our results produced a set of rankings of hospitals and local authorities according to their composite score with a 95% confidence interval reflecting uncertainty around this composite score. These results are illustrated in the left panel of figure 1. The dark dots show the composite score for hospitals arranged in order from worst to best—these dots assume all variation is due to genuine differences in performance. This is how most composite indicators are presented in practice—assuming there is no uncertainty at all. Around each of the dark dots the vertical line shows the 95% confidence interval arising from the simulations—these vertical lines assume all variation is random. This naive view of variation reflects a considerable degree of uncertainty in the composite score since the confidence intervals overlap over almost the entire range of performance. This means we cannot distinguish whether a hospital ranked 10 is genuinely performing better than a hospital ranked 50, because there is a great deal of uncertainty in capturing true performance variations—a hospital's score could fall anywhere along this vertical line. Similar results were obtained for local authorities.

However, after taking account of random variation in the underlying performance

Figure 2. Composite ranking and 95% confidence interval for 97 local authorities.



indicators, the results in the right panel of figure 1 illustrate that we were able to estimate genuine performance variations. It is now possible to say with more certainty that (for example) the hospitals at the bottom are scoring less well than those at the top of the league table, though there is still overlap in the middle. This is an improvement over the usual presentation of composite indicators as just the dark dots, where we assume all variation is due to differences in performance. With these confidence intervals taking uncertainty into account, and allowing for random variation, we can make much more robust statements about differential performance.

Similarly, when examining the rankings rather than scores of organizations, we obtain the results in figure 2 for local authorities. Authorities are ranked from worst (97th) to best (1st) and the vertical lines show the 95% confidence intervals around these rankings. Again, the left panel illustrates the high degree of uncertainty in the rankings of authorities prior to taking account of random variation in the underlying performance indicators with almost all confidence intervals overlapping. The right panel illustrates that this naive view, attributing all variation to randomness, is radically altered (with still some overlap in the middle) after accounting for random variation. Results are similar when the analysis is repeated for hospitals.

A similar pattern was found in the degree of random variation in the underlying individual indicators when comparing results across the two sectors. For healthcare, it varied from 80% (inpatient waiting times) to 1% (sickness absence rates). For local authorities, it ranged from 98% (unauthorized absences

secondary schools—education) to 1% (older people helped to live at home—social services). The full set of results is shown in table 1. There is a similarly wide range in the estimated proportion of variation across the 10 hospital performance indicators and the 35 local authority performance indicators. Managers will therefore have varying degrees of control over these performance indicators. Some of the random variation may be driven by slight changes in definition to the indicators over time, subtle changes in the way the data is collected or measured over time, performance targets attached to individual indicators which may lead to increased variation within organizations over time as they improve their performance, and possible ‘gaming’ behaviour.

#### *Examining Alternative Aggregation Methods*

Small changes in methods used to aggregate underlying indicators to construct the composite indicator can have a substantial impact on the results.

In the local government sector (under CPA), a differential weighting is applied to the various domains as set out in table 2. We explored the impact on the original composite indicator of changing the weights on the underlying performance indicators. Table 3 shows the impact of increasing and decreasing the weights on the performance indicators in education and social services by a factor of four and increasing and decreasing the weights on the performance indicators in environment and housing by a factor of two. The final column shows the impact of simultaneously amending the weights for the seven domains according to table 2. The results highlight the change in ranking across 97 places for the top, middle

**Table 1. Proportion of random variation on underlying performance indicators.**

<i>Variable</i>	<i>Proportion random variation</i>
<i>Healthcare</i>	
Percentage of patients waiting six months or less for an inpatient admission	80
Percentage of junior doctors complying with New Deal on Working Hours	66
Percentage of patients seen within 13 weeks of GP referral for first outpatient appointment	42
Summary measure of hospital episode statistics (HES) data quality	34
Inpatient survey satisfaction with co-ordination of care	32
Death within 30 days of surgery per 100,000 patients (non-elective admissions)	27
Emergency readmission to hospital within 28 days of discharge following hip fracture	25
Returning home within 28 days of emergency admission to hospital for fractured hip	9
Responses from NHS-employed staff survey on satisfaction with employer	2
Sickness absence rate—amount of time lost through absence as a percentage of staff time available	1
<i>Local government</i>	
Percentage of half days missed due to unauthorized total absences in secondary schools—education	98
Percentage of total tonnage of household waste which has been recycled—environment	60
Percentage of interactions with citizens delivered electronically—corporate health	58
Percentage of 15-year-old pupils with five or more GCSEs at grades A*–C—education	54
Percentage of working age people with disabilities as a percentage of economically active—corporate health	36
Percentage of total tonnage of household waste sent for composting—environment	31
Percentage of permanently excluded pupils attending alternative tuition: 20 hours or more a week—education	27
Percentage of major planning applications in 13 weeks—planning	25
Number of private dwellings six months empty: returned to occupation or demolished—housing	24
Community strategy developed in collaboration with local strategic partnership?—corporate health	22
Percentage of working age people from ethnic minorities as percentage of economically active—corporate health	22
Number of recorded domestic burglaries per 1,000 households—community safety	19
Average gross weekly cost of intensive social care for adults and older people—social services	18
Housing benefit security—number of fraud investigations per 1,000 caseload—benefits	18
Percentage of care leavers in education/training/employment—social services	15
Percentage of employees retiring on grounds of ill-health as percentage of total staff—corporate health	14
Score on creating opportunities checklist for adoption of local cultural strategy—culture and libraries	13
Condition classified non-principal roads by coarse visual inspection (CVI) survey—transport	12
Percentage of new homes built on brown field sites—planning	8
Percentage of people receiving needs statement and how they will be met—social services	8
Percentage of top 5% of earners in authority that are women—corporate health	7
Percentage of standard searches carried out in 10 working days—planning	6
Percentage of racial incidents that resulted in further action—community safety	5
Number of road accident casualties per 100,000 population: car users—transport	4
Percentage of statements of special educational need (SENs) issued in 18 weeks with exceptions—education	4
Percentage of total length of footpaths easy to use by the public (Country Agency and CSS Survey)—transport	3
Percentage of principal roads not needing major repair / average structural expenditure per km—transport	3
Number of kilograms of household waste collected per head—environment	2
Percentage of renewal claims processed on time—benefits	2
Number of visits to libraries per 1,000 population—culture and libraries	2
Average length of stay of households in hostels—housing	1
Percentage of children looked after with three or more placements—social services	1
Percentage of schools identified by Ofsted as requiring special measures—education	1
Percentage of secondary schools with 25% or more places unfilled and at least 30 surplus places—education	1
Older people helped to live at home per 1,000 population aged 65 or over—social services	1

and bottom five local authorities. The correlation between the new and original rankings varies between 0.81 and 0.96. The largest jump in position for an individual authority was 54 places, more than half the league table. On average, authorities changed between six and 13 places in the rankings, depending on the changes made to the weighting system. Clearly, changes to the weighting structure of performance indicators can have a profound impact on the rankings of organizations. Results were again very similar when the analysis was repeated for hospitals.

We also illustrate that using decision rules to assign hospitals or local authorities to ordinal

categories (for example 0–3 stars or ‘excellent’, ‘good’, ‘poor’) which are typically applied in the construction of composite scores, produces even more variability in ratings.

**Table 2. Weighting applied in the CPA.**

<i>Seven domains</i>	<i>Weight</i>
Education	4
Social services	4
Environment	2
Housing	2
Libraries and leisure	1
Benefits	1
Use of resources	1

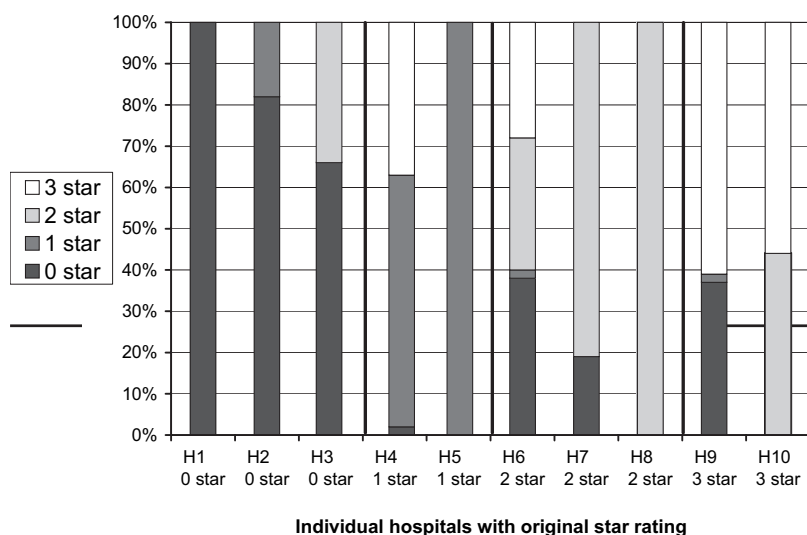
**Table 3. Rankings for local authorities after changing weights on underlying performance indicators.**

	<i>Original</i>	<i>(Education + social services) × 4</i>	<i>(Education + social services) × 1/4</i>	<i>(Environment + housing) × 2</i>	<i>Environment + housing) × 1/2</i>	<i>(Education + social services) × 4 + (environment + housing) × 2 + rest × 1</i>
Top 5	1 2 3 4 5	3 21 2 5 30	4 1 13 7 3	1 2 11 5 8	9 5 1 4 7	9 10 4 3 33
Middle 5	47 48 49 50 51	42 58 55 54 86	62 43 52 54 28	35 51 30 62 31	57 49 64 47 67	35 75 60 55 70
Bottom 5	93 94 95 96 97	84 79 43 97 95	91 93 96 76 97	90 91 94 84 97	95 94 93 96 97	90 93 65 96 97
<i>Correlations</i>		0.81	0.88	0.91	0.96	0.88
<i>Largest change</i>		52	54	42	23	38
<i>Average change</i>		13	10	9	6	11

Results in figure 3 show the high degree of uncertainty which is introduced by the use of these decision rules. The figure illustrates the proportion of times a hospital is placed in each of the four star categories, using a sample of 10 hospitals, randomly selected from each of the 4 categories (based on the original assignment to star category). We found that while there was

relative stability for the worst performing hospitals (with zero score), this was not the case for the best performers (three stars). So, for example, hospitals H1 and H2 achieved a zero score in 100% and 82% of the simulations respectively; whereas hospitals H9 and H10 received the score of 3 stars only 61% and 56% of the time respectively. Indeed, because of the threshold nature of the decision rules, hospital H9 suffers a catastrophic relegation to a zero score in 38% of the simulations.

Table 4 summarises the results for all 117 hospitals and reiterates the greater stability in the ranking of the worst hospitals over the 1,000 simulations. Similar results were obtained for local authorities.

**Figure 3. Proportion of times a sample of 10 hospitals receive a particular rating on the composite constructed from decision rules.**

### Conclusions

The production and publication of performance league tables in public services in England has most often been linked to a regulatory regime with a set of sharp incentives intended to change behaviour and drive performance improvements. Organizations are rewarded or punished according to the outcome of the league table. Yet there has been a paucity of research on how these composite performance indicators are constructed, what the methodological challenges are in doing so, and whether they are in fact a good reflection of performance. Despite this lack of evidence, their widespread appeal to policy-makers seems

**Table 4. Proportion of times hospitals receive a particular rating on the composite constructed from decision rules.**

No. of hospitals	New composite category	Percentage of times in simulations that composite is given a score of:			
		0	1	2	3
33	0	79.0	8.6	11.9	0.4
22	1	7.7	77.4	8.2	6.7
46	2	9.8	10.0	75.5	4.7
16	3	9.1	17.7	19.5	53.7

set to continue. In light of their likely continued use by policy-makers, we sought to contribute to a better evidence and understanding of the technical challenges in constructing composite performance measures so that their impact on behaviour might be better anticipated and managed. The use and publication of composite performance measures can generate both positive and negative behavioural responses and if significant policy and practice decisions rest on the outcome of the composite, it is important to have a clear understanding of the potential risks involved in constructing a composite and arriving at a ranking.

We asked whether performance measurement based on composite indicators is robust. We found a considerable degree of uncertainty in the construction of composite indicators with little guarantee that any consistent league table ranking can be secured. This concurs with previous research that suggests that rankings of performance indicators are unstable (Goldstein and Spiegelhalter, 1996; Marshall and Spiegelhalter, 1998). The construction of composite indicators is sensitive to methodological choices. Changes in aggregation methods (either altering weightings or decision rules) can have a substantial impact on results, with organizations jumping from one end of the league table to the other dependent on relatively small alterations in the aggregation rules.

We also asked to what degree composites are influenced by the uncertainty surrounding the underlying indicators on which they are based. In any performance benchmarking system, we need to know an estimate of the degree of random variation for each indicator so that we can draw definitive conclusions about real differences in performance. When we take account of random variation, we gain much greater precision in performance assessment. We disentangled genuine performance variations (for which managers can be held accountable) from random fluctuation in the measurement of performance

indicators. We found large differences across different types of indicators on the estimate of variation which is within management control. Stripping out the variation for which managers can be held accountable drastically reduced the amount of uncertainty in the league table and produced much more robust results.

Key implications for policy and practice are:

- ‘Decision rules’ need to be treated with caution. Subtle and highly subjective changes to the decision rules can dramatically impact on the composite index and rankings of organizations.
- The choice of a weighting system can have a significant impact on the rankings of individual units within the composite. The choice of weights may be ad hoc and arbitrary with a lack of consideration for whose preferences the weights reflect and how robust these are. Greater attention should be paid to the origin and nature of weights and the sensitivity of composites to changes in the weighting structure.
- The proper treatment of uncertainty in composite performance measures is crucial—composites need to be published with indications of uncertainty to communicate the sensitivity of the reported measure. ■

#### Acknowledgements

This research was funded by the Economic and Social Research Council (ESRC) grant number RES-153-25-0031 under the Public Services Programme. The views expressed are those of the authors and not the funders. We should like to thank participants at numerous conferences for helpful comments on earlier drafts of this article including the European Conference on Health Economics (Budapest); the Public Services Workshop Series on Ranking Public Services (Oxford); the National Institute of Economic and Social Research Public Sector Performance Conference (London); and the European

Institute for Advanced Studies in Management Conference (Nice). Any errors which remain are solely those of the authors.

### References

- Audit Commission (2005), *Service Assessment Framework: Technical Guide to CPA 2005 for Single Tier and County Councils* (London).
- Dolan, P., Gudex, C., Kind, P. and Williams, A. (1996), Valuing health states: A comparison of methods. *Journal of Health Economics*, 15, pp. 209–231.
- Freudenberg, M. (2003), *Composite Indicators of Country Performance: A Critical Assessment*, OECD STI Working paper DSTI/DOC 2003/16 (OECD, Paris).
- Goldstein, H. and Spiegelhalter, D. J. (1996), League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159, 3, pp. 385–443.
- Healthcare Commission (2005), *2005 Performance Ratings* (London).
- Hood, C. (2007), Public service management by numbers: why does it vary? Where has it come from? What are the gaps and the puzzles? *Public Money & Management*, 27, 2.
- Joint Research Centre (2002), *State of the Art Report on Current Methodologies and Practices for Composite Indicator Development* (Applied Statistics Group, Institute for the Protection and Security of the Citizen, European Commission).
- Marshall, E. C. and Spiegelhalter, D. J. (1998), Reliability of league tables of *in vitro* fertilization clinics: retrospective analysis of live birth rates. *British Medical Journal*, 6, 316, pp. 1701–1705.
- Mooney, C. Z. (1997), *Monte Carlo Simulation*, No. 116 in series on quantitative applications in the social sciences (Sage Publications, London).
- Mullen, P. and Spurgeon, P. (2000), *Priority Setting and the Public* (Radcliffe Medical Press, Abingdon).
- Smith, P. (2002), Developing composite indicators for assessing health system efficiency. In Smith, P. C. (Ed), *Measuring Up: Improving the Performance of Health Systems in OECD Countries* (OECD, Paris).
- Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data* (The MIT Press, London).

# The Perils and Pitfalls of Performance Measurement: The CPA Regime for Local Authorities in England

**Iain McLean, Dirk Haubrich and Roxana Gutiérrez-Romero**

*From 2002 comprehensive performance assessment (CPA) has been used by the Audit Commission to scrutinize service delivery in English local authorities across six service blocks: benefits; social care; environment; libraries and leisure; use of resources; education and housing. The authors examined CPA in terms of how vulnerable it is to categorization errors and gaming, whether it is consistent with other government policies and how it deals with uncontrollable factors. CPA failed all of these tests.*

As with many other parts of the public sector, recent UK governments of both main parties have shown increasing interest in using performance metrics as a tool to assess and improve the performance of local authorities. The importance of a well-designed instrument for measurement and improvement cannot be exaggerated: British local authorities are responsible for about a quarter of all public spending in the UK, of the order of £120 billion per annum, raise only about a quarter of that from local taxation, and rely for the remainder on transfer payments from central government (ODPM, 2005, p. 21). In 2001, the government embarked on an elaborate evaluation exercise—comprehensive performance assessment (CPA)—to gauge how effectively its money was being spent to provide high-quality local services (DTLR, 2001; Audit Commission, 2007).

From 2002 CPA has been used to scrutinize service delivery in English local authorities across six service blocks: benefits; social care; environment; libraries and leisure; use of resources; education and housing. Hundreds of performance indicators and a plethora of audit and inspection reports have been collected, summarized, weighted, and categorized so as to arrive at final category ratings of 'excellent', 'good', 'fair', 'weak', and 'poor'.

During the first three annual rounds, the CPA exercise was applied to the 148 English unitary and upper-tier authorities only; it was extended in 2005 to include also the 238 District authorities. The analysis in this article applies

to the regime that was in place from 2002 to 2004. Some methodological changes were introduced in the 2005 CPA round, and were too late for our analysis. However, these changes do not appear to have altered the basic premises of the regime, so the analysis here is valid for the current framework.

The 'carrots' offered to authorities for good performance consisted of exemptions from future inspections, the removal of council tax caps, and less ring-fencing of central grants. These rewards are collectively known as 'earned autonomy'. Conversely, authorities that failed to perform incurred tighter and more frequent inspections, which could go as far as the imposition of external monitoring boards that would meet at three-monthly intervals and scrutinize the authority's progress on a continuous basis.

The results of the CPA regime during our study period were as follows:

- The number of 'poor' and 'weak' councils dropped from 34 in 2002 to 16 in 2004.
- The number of 'weak' councils in 2004 dropped to 15, down from 21 in 2002.
- Finally, 101 councils achieved a rating of 'excellent' or 'good' in 2004, up from 76 in 2002.

But did the improvements in CPA scores reflect real improvements in authorities' performance? Does CPA measure the right 'things'? Is it possible for local authorities to improve their CPA scores without improving

*Iain McLean is Professor of Politics, University of Oxford.*

*Dirk Haubrich is a research officer in the Department of Politics and International Relations, University of Oxford.*

*Roxana Gutiérrez-Romero is a research officer in the Department of Social Policy and Social Work, University of Oxford.*

**Table 1. The conversion of performance assessment scores to category scores.**

<i>Performance score (shire counties)</i>	<i>Category score</i>	<i>Performance score (London boroughs, and metropolitan and unitary councils)</i>	<i>Category score</i>
Less than 24 points	1	Less than 30 points	1
24 to 29	2	30 to 37	2
30 to 36	3	38 to 45	3
More than 36	4	More than 45	4

performance? In short, is CPA a reliable and valid way of assessing performance?

The CPA framework consisted of two separate assessments that evaluate the authorities' current performance and their ability to improve in the future.

### Current Performance

Current performance of authorities was assessed in seven categories (benefits; social care; environment; libraries and leisure; use of resources; education and housing). Where available, performance was assessed through already existing judgments from inspectorates and auditors, such as those by Office for Standards in Education (Ofsted) and Department for Education and Skills (DfES) for the education service block. Numerous categorizations and conversions were applied in order to summarize more than 1,000 performance indicators and auditor judgements. Eventually, authorities obtained a score between 1 and 4 for each of the service blocks (with 1 being the lowest, and 4 the highest). The scores were then weighted so that the scores for education and social services count four times, housing and environmental services twice, with the remaining blocks counting only once. These were then added up to produce a performance score of between 15 and 60 points, or 12 and 48 points for shire county councils (because they do not provide, and are therefore not assessed on, housing or benefits services). The performance scores were then categorized to produce a performance rating of between 1 and 4 for each authority (see table 1).

A second assessment concentrated on a council's plan to improve services in the future, which itself consisted of two components:

- A self-assessment.
- An external 'corporate assessment' carried out by a small team of auditors, inspectors, officers, and members from peer councils, all of whom were tasked with checking the

statements made by authorities in the self-assessment.

A total of nine areas were assessed that way, each of were given a score between 1 and 4. These were then weighted to produce an overall corporate assessment score ranging between 12 and 48 points. This score was then converted from continuous data to categorical data, to give a score between 1 and 4 (see table 2).

The two assessment ratings (on current performance and ability to improve) were then combined to produce an overall assessment: 'excellent', 'good', 'fair', 'weak' or 'poor'. Final CPA ratings were not, however, a product of simple arithmetic. Additional minimum thresholds are applied in the calculations to account for cases where one of the council's ratings significantly deviated from the other. Table 3 provides an overview of the resulting score matrix that determined councils' eventual CPA rating.

We carried out a statistical analysis of the first three CPA rounds that were conducted for the 148 English unitary and upper-tier authorities in the years 2002, 2003, and 2004. Additional insights and pointers are provided from semi-structured elite interviews.

The CPA regime has been in operation in this form since 2002, with minor alterations introduced in the most recent 2005 round, when the CPA ratings were replaced by NHS-style star ratings and self-assessments were

**Table 2. The conversion of CPA's 'ability to improve' scores to category scores.**

<i>Council score</i>	<i>Category score</i>
Less than 24 points	1
24 to 31	2
32 to 39	3
More than 40	4

replaced by 'direction of travel statements' (which were still based on self-assessments). CPA is a typical example of a 'top-down performance management' regime as stipulated by the government's approach to public service reform, which identifies market incentives; capability and capacity improvement; users shaping services from the bottom up; and performance management from the top as the four core elements through which public service delivery is to be modified across all service areas (see figure 1).

In the introduction to this issue, Christopher Hood (2007) identifies three possible applications for performance measurement of public services:

- Targets.
- Rankings.
- Wallpapers.

CPA generated targets that stipulate floor standards and thresholds to determine an authority's category score, either on a particular indicator, a service block, or the overall performance of the authority; and rankings that allow comparisons between authorities, again per indicator, service block or overall performance.

It should probably come as no surprise, then, that the CPA regime was marked by the deficiencies that Hood attributes to these two applications, namely categorization errors, gaming and, to add two of our own, contradictory incentives and external constraints. We take on each in turn.

#### Categorization Errors

The numerous conversions illustrated in the tables indicate the extent to which the assessment framework was vulnerable to categorization errors: the rule-based categorizations that lead to the final CPA ratings in table 3 are as arbitrary as the category thresholds used in tables 1 and 2 to arrive at the performance and improvement scores respectively. Moreover, the categorizations used were chosen *after* the performance data were collected in the initial 2002 CPA round, so it is possible that the categorizations themselves were selected to best serve a specific purpose.

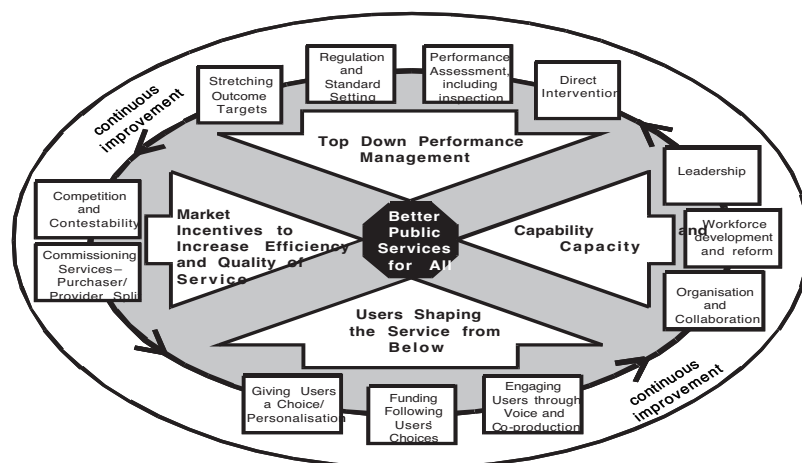
It may, therefore, not have been a coincidence that in the initial CPA round just over half of the authorities (76) ended up in the top two categories of 'excellent' or 'good' (Audit Commission 2002a; 2002b). For in so doing, the poor performers could be easily treated as a 'minority' in need of guidance and further

inspection in order for them to move up to the 'majority' of good performing councils. Had the distributional split between goodies and baddies resulted in the latter outnumbering the former, the newly-imposed CPA regime could have been accused of being unrealistic, would have faced greater political resistance as a result, and would have most likely failed.

Such categorization errors may not necessarily compromise the reliability of the CPA regime (it was consistent in assessing the units under observation) but certainly its validity (because, although consistent, it was consistently off-target). The 'ability to improve' component of the CPA regime allows for some linear regression analysis to be carried out that measures the extent to which CPA has been off target, because the 'ability to improve' scores in the 2002 CPA round can be correlated with the actual improvement achieved in 2003 and 2004. The analysis shows that the former exhibits no correlation with the latter. In other words, ability to improve scores are a very bad indicator for actual improvement in subsequent years (see Haubrich and McLean, 2006b, for further details). Thus the 2002 self-assessment scores were either categorization errors (i.e. they wrongly measured authorities' ability to improve), or gaming errors (i.e. authorities regarded them as 'cheap talk' and therefore an easy way to boost their CPA scores).

This is a disturbing conclusion, even more so because the Audit Commission is deliberating, at the time of writing this article in March 2007, how best to re-conceptualize CPA for the rounds 2008 onwards. One of the ideas discussed is substituting for CPA a less

**Figure 1. The UK government's approach to public service reform (Prime Minister's Strategy Unit, 2006, p. 8).**



costly regime that sees authorities assessing their own performance. It may also carry lessons for healthcare performance managers, where practice is changing rapidly (see the article by Jacobs and Goddard in this issue; Healthcare Commission, 2007)

Improvements in the methodology of self-assessments is required, then, if the assessment of local authorities is to have any meaning at all. Wales and Scotland may lead the way on this topic, as both nations grant their authorities considerable flexibility, including self-assessments. Yet, even they could not escape the conundrum that self-assessment leads to heterogeneity in assessment priorities, assessment methods and assessment reporting, which eventually makes the assessment results no longer comparable across the inspected units. League tables of indicators or composite ratings, which are neither published nor compiled in either Wales or Scotland, would have to be scrapped in England as well.

### Gaming

Some of the theoretical problems of performance metrics are well known (Boyne, 2002). Their most general formulation is Goodhart's Law. In its original formulation, 'Any observed statistical regularity will tend to collapse once pressure is placed on it for control purposes' (Goodhart, 1984, p. 94). More pithily, 'when a measure becomes a target, it ceases to be a valid measure'. Performance indicators are subject to gaming, in the sense of hitting the target but missing the point, or reducing performance where targets do not apply (Bevan and Hood, 2006, p. 521). In his introduction to this issue, Hood (2007) identifies three forms of gaming—threshold effects, ratchet effects, and output distortions—at least two of which emerged during the elite interviews we conducted with chief executives and directors of performance in local authorities. Although it is difficult in such interviews to ascertain any outright admission of cheating, or even gaming, we did manage to identify several output distortions and threshold effects.

#### *Output Distortions*

Output distortions are attempts to achieve targets at the cost of significant but unmeasured aspects of performance. We detected gaming of this kind with an output indicator that is used to measure the extent to which an authority succeeds in promoting its leisure facilities to the public, namely the number of swims per square foot of pool area. The effort can easily be gamed in either of two ways:

- Closing some (or all) of an authority's pools except one (which all the die-hard swimmers would eventually have to be content with).
- Allocating pool slots to members of swim clubs (whose lane-swimming discipline allows for greater usage of each lane), at the expense of swimming lessons or general sessions for the public (which would require considerably more pool space per activity).

In both cases, the public would experience some deterioration in service quality, which should be reflected accordingly in the metric measuring this performance. Yet in both cases the performance indicator actually used would go up. In earlier work, Boyne (2002) had concluded that performance indicators had improved their validity between their first introduction in 1994 and the time of his conducting the research, yet our own research, and that of others (for example Wilson, 2004, table 1), indicates that much work remains to be done.

#### *Threshold Effects*

Threshold effects refer to the effects of targets on the distribution of performance among a range of, and within, the units assessed, 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, and more generally to crowd performance towards the target. Our research did not uncover cases that would confirm the penalization hypothesis of threshold effects, probably because the CPA regime does not stipulate a single target but imposes four layered targets (in the form of the four category scores) and does so not only for the final service block ratings (see tables 1 and 2) but also for each of the thousand indicators (not shown). There is therefore less scope for exceptional performers not to be rewarded.

However, we did find evidence of crowding towards the target, in ways that cannot be described as enhancing the performance of the authority. One interviewee reported that in his work as performance manager of a county council, he prioritizes 'quick-win' activities. These are measures that make the council move up a category on a particular indicator without incurring too much effort. This approach is possible because the CPA regime stipulates for many indicators not only a numerical value of performance to be achieved, but also the types and format of documentation that need to be produced or processes to be

shown in place. Improving the ranking therefore becomes possible by simple (sometimes one-off) administrative adjustments, without improving the underlying performance, both in itself and as measured by the performance indicator.

### Contradictory Incentives

Individual departments of government are often not joined up, leading to situations where the left hand of government is unaware of what the right hand is doing. In the case of the CPA regime we found one instance where one measurement regime contradicts and defeats another. Half of the CPA score was derived from an authority's current performance assessed across six service blocks. One of the most important and costly service blocks to the public is education. Just as in any of the other service blocks, an authority's CPA score in education was an increasing function of its performance on a number of indicators. One of these indicators was 'percentage of pupils in the authority doing well in key stages 2, 3 and 4'.

However, the same indicator was used in a different context (the funding of local authorities) and with a different aim (to measure an authority's level of deprivation). Given that local authorities with high levels of deprivation in their populations are faced with additional resource needs, central government grants should reflect deprivation levels accordingly. On an education-specific level, authorities obtain additional funds to meet special educational needs (SEN) and additional educational needs (AEN). For those additional grants, authorities themselves may determine which indicator to use as a proxy, and they have adopted a range of approaches as a result: free school meals, low pupil attainment rates, housing benefits, number of vulnerable children, deprivation, or a combination of these.

When deprivation is employed as an indicator, it is usually measured by the index of multiple deprivation (IMD) 2004. The IMD is a combined indicator calculated by academic researchers, at the request of the Department for Communities and Local Government (DCLG), and is intended to measure the relative deprivation across 32,482 'super output areas' in England. It measures 37 different indicators across seven domains: income; employment; health and disability; education, skills and training; barriers to housing and services; living environment and crime. For each domain, each of the output areas in England obtains a score and a rank. Results for each domain for

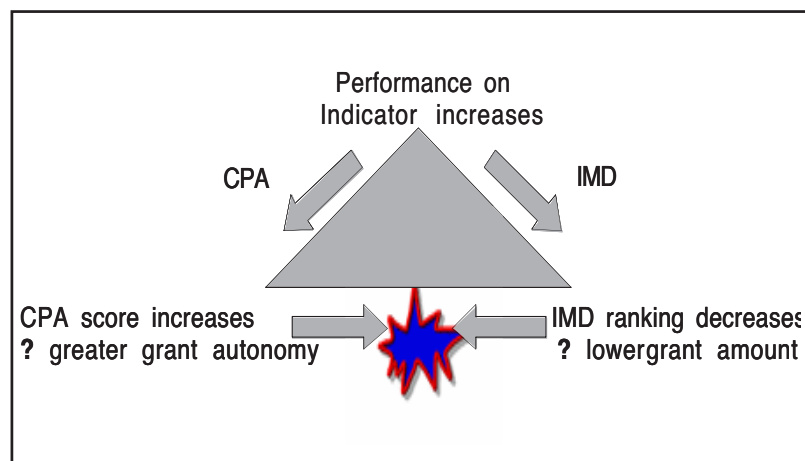
each area are then combined, with different weightings for each domain, to produce an overall IMD score and rank.

The IMD is widely recognized and standard across authority boundaries. Since its update from its less detailed predecessor IMD 2000, the IMD 2004 has been used by an increasing number of local authorities to calculate school funds and other personalization funds (DFES, 2006, p. 20). Central government has also used it in other national and local government contexts, to inform allocations of the neighbourhood renewal fund and, more generally, the formula spending share (FSS, formerly the standard spending assessment), which is the main determinant of an authority's revenue support grant (RSG). The FSS makes use of information reflecting the demographic, physical and social characteristics (including deprivation) of each area and provides various 'top-ups' to the basic grant (ODPM, 2005).

One of the indicators used in the IMD 2004 is 'percentage of pupils in the authority doing well in key stages 2, 3 and 4', which means that an authority's IMD score depends on the authority's performance on this and 26 other indicators. Yet, as noted at the beginning of this section, an authority's CPA score was informed by its performance on the very same indicator. A contradictory incentive mechanism emerges (see figure 2) to which authorities are exposed: an authority's IMD score in the education domain is a *decreasing* function of its performance on the above indicator, whereas its CPA score in education was an *increasing* function of its performance.

The implications of this are significant because a chief executive of an authority faces the choice between improving the authority's

**Figure 2. The vicious triangle of measuring educational performance.**



schools and obtaining a higher CPA score (which results in greater 'earned autonomy' of deciding how best to spend central grants) but a lower IMD score (which results in a reduction in the amount of the grant), or letting them deteriorate (and obtaining a higher grant but less autonomy to spend it). Either way, the authority gains a financial bonus and suffers a financial penalty and should do whichever carries the larger net bonus.

Government policy is contradictory here: central government is rewarding local authorities in poor areas for their deprivation, and punishing them for their poor CPA scores. And rich areas are being rewarded for their good CPA scores and punished for their lack of deprivation. If earned autonomy costs less and/or allows the authority to raise money more cheaply than unearned non-autonomy, government policy comes dangerously close to being self-defeating. The implication for central government is that there are two contradictory regimes in place, at least one of which must be abandoned or modified. Similar issues have arisen in services other than those provided by local government, such as healthcare, where primary care trusts with high levels of poverty in their areas received poor star ratings because Ministers refused to standardize for deprivation, on the grounds that that would perpetuate inequalities (Bevan, 2006).

### External Constraints

To be a valid measure, a performance score should be entirely uncorrelated with any background factor that a local authority can neither control nor be blamed or praised for. At least in the short run, an authority cannot control the ethnic mix, age profile, morbidity or mortality of its local population. Many of these factors are bundled into indices of deprivation. The Audit Commission has maintained that it isn't any harder (or easier) for an authority in a deprived area to score highly in CPA than for an authority in a prosperous area. The adjustments the Audit Commission introduced in the 2005 CPA round were published too late to be included in our analysis. Further, academic commentators have always been sceptical, as to which see an earlier special issue of this journal (*Public Money & Management*, Vol. 24, No. 1, January 2004). Our analysis leads us to side with the sceptics.

On individual services, there is reason for the unwanted correlation to run in either direction. For instance, if authorities in deprived areas get extra resources, they may deliver better services. Alternatively, authorities in

deprived areas may find it harder to get a high CPA score. This may be tautologically true, as we have just shown in the case of school performance. Or it may be contingently true, for instance if it is harder to get a good score when teaching (or providing social services to) multi-ethnic populations than to mono-ethnic populations.

For performance scores to be uncorrelated with deprivation, the extra grant to deprived areas would need to exactly compensate for their deprivation. This is a tall order. It is not surprising that it remains unmet. We have analysed the relationship between CPA score and deprivation in finer detail than previous analysts and we found that:

- The CPA scores from 2002 to 2004 were positively affected by discretionary expenditures incurred by local authorities, so authorities can 'buy' better scores by spending beyond the level regarded as appropriate by central government.
- When grouping all local authorities together, deprivation in the domains *education*, *crime* and *living environment* had a significant negative effect on overall CPA score, in other words, the higher the deprivation in these domains, the lower the CPA score obtained. However, these effects can vary by type of local authority and CPA domain, which reflects the fact that the degree of deprivation, too, varies across type of authority.

### Conclusion

The CPA regime in place in England between 2002 and 2004 would be a reliable and valid measure of local authority performance if it passed the tests to which we have subjected it:

- Invulnerable to categorization errors.
- Invulnerable to gaming.
- Consistent with other government policies.
- Uncorrelated with uncontrollable factors.

We have shown that it fails all four tests. As governments are likely to continue using performance measures, it is important that, in 2006 jargon, they be fit for purpose. Those for English local authorities have some way to go. ■

### Acknowledgements

The authors would like to thank the ESRC for funding, through project grant RES-153-25-00 60, the research that led to this article, as well as the two anonymous reviewers for

providing helpful comments. We also thank Michael Noble and David McLennan for giving access to the detailed IMD 2004 data and for helpful comments.

## References

- Audit Commission (2002a), *A Picture of Performance: Early Lessons from Comprehensive Performance Assessment* (London).
- Audit Commission (2002b), *Comprehensive Performance Assessment—Scores and Analysis of Performance* (London).
- Audit Commission (2007), *Why CPA?* (London).
- Bevan, 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, G. and Hood, C. (2006), What's measured is what matters: targets and gaming in the English public health care system. *Public Administration*, 84, 3, pp. 517–538.
- Boyne, G. A. (2002), Concepts and indicators of local authority performance: an evaluation of the statutory frameworks in England and Wales. *Public Money & Management*, 22, 2, pp. 17–24.
- DfES (2006), *Technical Review of Deprivation Indicators (September 06)*. Available from [www.teachernet.gov.uk](http://www.teachernet.gov.uk)
- DTLR (2001), *Strong Local Leadership—Quality Public Services*, CM 5237 (London).
- Goodhart, C. A. E. (1984), *Monetary Theory and Practice. The UK Experience* (Macmillan, London).
- Haubrich, D. and McLean, I. (2006a), Assessing public service performance in local authorities through CPA—a research note on deprivation. *National Institute Economic Review*, 197, 1, pp. 93–105.
- Haubrich, D. and McLean, I. (2006b), Evaluating the performance of local government: a comparison of the assessment regimes in England, Scotland and Wales. *Policy Studies*, 27, 4, pp. 272–293.
- Healthcare Commission (2007), *The Annual Health check in 2006/2007: Assessing and Rating the NHS*. Available from [www.healthcarecommission.org.uk](http://www.healthcarecommission.org.uk)
- Hood, C. (2007), Introduction. *Public Money & Management*, 27, 2.
- Jacobs, R. and Goddard, M. (2007), How do performance indicators add up? An examination of composite indicators in public services. *Public Money & Management*, 27, 2.
- ODPM (2004), *The Formula Grant Distribution System—Supporting Key Public Services*. Available from [www.local.odpm.gov.uk](http://www.local.odpm.gov.uk).
- ODPM (2005), *Local Government Finance Statistics England No. 16* (London).
- Prime Minister's Strategy Unit (2006), *The Government's Approach to Public Service Reform* (Cabinet Office, London).
- Wilson, J. (2004), Comprehensive performance assessment—springboard or dead-weight? *Public Money & Management*, 24, 1, pp. 63–68.

## *Public Money & Management:* Instructions for authors

*Public Money & Management* (PMM) publishes articles which contribute new knowledge as a basis for policy or management improvements, or which reflect on evidence from public service management and finance. The journal does not accept literature reviews. PMM has a multidisciplinary readership, including officials in all types of public service organizations, academics, consultants and advisers working with the public services, politicians, journalists, and students on both academic and professional courses. Although this readership has been interested largely in British public management, there is an increasing interest in international developments. Accordingly, the editors welcome articles about developments outside the UK which offer clear lessons for British or other western practitioners.

The journal publishes main articles, new developments and contributions to debate. *Main articles* (no more than 5000 words) must meet high standards of intellectual argument, evidence and understanding of practice in public management. They are double-blind refereed by both an academic and a practitioner. *New developments* (up to 2750 words) focus on the evolution of contemporary public service policy, management or practice and convey the potential or actual impact of change in a detached, informed and authoritative way. These articles are not normally refereed, but are subject to editorial scrutiny. *Debate* articles (usually under 1000 words) are personal statements about topical issues, expressing an argument, supported by examples or evidence. They, too, are subject to editorial scrutiny. Authors should take into account the needs of the readership in drafting their articles and, in particular, to explain technical terms and avoid exclusive jargon.

PMM is published six times a year in January, March, May, July, September and November.

### Submission

Manuscripts should be emailed to the managing editor, Michaela Lavender:  
**michaela.lavender@cipfa.org.uk**

### Preparation

The following items should be included with manuscripts: title; full postal and email addresses of all authors; a one-line biography about each author; and PDFs of any illustrations in black and white to fit our A4 page. Spelling should follow the *Oxford English Dictionary*. Authors should also supply a clear summary of around 50 words focusing on conclusions and lessons.

The journal uses the Harvard (author, date) system of referencing. References in the text should be given as (Brown, 1990), or Brown and Jones (1990), or Brown *et al.* (1992) if there are three or more authors. References should be given at the end of the article in a single alphabetical list:

*To a journal:* Jones, G. and Stewart, J. (2009), Accountability in public partnerships—the case of LSPs. *Public Money & Management*, 29, 1, pp. 59–64.

*To a book:* Parker, D. (2009), *The Official History of Privatization* (Routledge, London).

Documents available only online should be listed as: 'author/editor (year), Title. See [URL]'. References to personal emails and private correspondence should be avoided. Footnotes and endnotes should also be avoided if possible.

### Final manuscripts and proofs

After acceptance, authors are requested to submit their final manuscripts by email to the managing editor: michaela.lavender@cipfa.org.uk. Proofs for checking will be sent to authors by email and should be returned promptly by fax, or emailed as a marked-up and scanned PDF, to Michaela Lavender on 001 561 989 9968.

**For more information about PMM, see <http://www.cipfa.org.uk/pt/pmm>**

# Performance, Strategy and Accounting in Local Government and Higher Education in the UK

**Martin Broad, Andrew Goddard and Larissa Von Alberti**

*This article discusses the importance of organizational management of performance measures. The authors use a grounded theory methodology to explore the relationship between strategic planning, accounting and performance measurement systems in local government and higher education. Only by understanding how and why performance management works will it be possible to improve the delivery of our public services.*

It is not only the technical aspects of performance measures which are important in understanding their spread and increasing influence through the public sector, but also the way they are managed by organizations. Consequently, this article reports an investigation of the organizational management of performance measures (PMs) in public service cases in the south of England. There is a good deal of experience of using PMs in innovative ways in both the private and public sectors. In the private sector, approaches such as the balanced scorecard (BSC) have emerged as ways of managing performance. The BSC was developed by Kaplan and Norton (1992) as a tool to track the key elements of a company's strategy by developing and monitoring a set of PMs. Kaplan and Norton (1992) stress that the BSC puts strategy and vision, and not control as its emphasis. The research undertaken in these developments suggests that both strengths and weaknesses occur in practice. A good deal of research has also been undertaken in the area of strategic management accounting (SMA). A particular feature of SMA is the combination of financial and non-financial measures to enhance organizational performance. Researchers recognize the importance of both sets of measures and of understanding the relationships between them. Financial measures are seen as inadequate measures of long-term strategic performance and the search for more comprehensive performance measurement systems has resulted.

Different initiatives have also been undertaken in performance management in different parts of the public sector. For instance, the use of performance league tables has

emerged in local government and health, best value initiatives in local government, BSC and strategic costing approaches in higher education and quality of life and other output measurement approaches in health. However, there has been relatively little empirical research in the public sector into the use of such measures. This article addresses some of the gaps in empirical public sector performance management research.

The public sector comprises a broad range of organizations and services and it is not feasible to include all of these in one study. The research reported here was an exploratory, pilot study designed to explore the issues in a limited part of the public sector. Local government and higher education were selected because they possess a range of the characteristics found across the sector which might be expected to influence performance management.

The first concerns the different governance structures. Local government is characterized by directly-elected councillors who are ultimately responsible for the management of their organizations. Councillors have to stand for election on a regular basis and are therefore held to account in a very public manner. There are no directly-elected members in higher education.

The second concerns the broad range of services provided by local government compared to the narrow range provided by higher education. The organizations included in this study were 'unitary' authorities which provide all the local government services within a defined geographical area, for example schools, social services, housing, recreation and culture, refuse collection and disposal, highways

*Martin Broad is a lecturer in management accounting in the School of Management, University of Southampton.*

*Andrew Goddard is professor of accounting in the School of Management, University of Southampton.*

*Larissa Von Alberti is a doctoral student in the School of Management, University of Southampton.*

and transportation, town and country planning and economic development. Higher education, of course, is more restricted, although it does include a range of disciplines across teaching and research.

The third difference between the sectors concerns the extent to which the organizations are exposed to market mechanisms. Higher education is characterized by a quasi-market for undergraduates and a real market for post-graduates. Its income is therefore significantly affected by these markets. Local government income is not as exposed to markets because the vast majority of its income is raised through direct tax and government grant. Each of these differences may be expected to result in different performance management in the two sectors and be indicative of the range which might be found across the whole public sector.

To explore the relationship between strategic planning, accounting and performance measurement systems, a grounded theory methodology was used. This comprised four case studies in local government and higher education. This methodology is inductive and emphasises the importance of allowing issues to emerge from the data. Rather than test a set of hypotheses, a set broad research questions are used to develop a nascent theory of the relationship.

### Prior Research

Empirical studies have investigated the success of the BSC in the private sector with mixed results. Hoque and James (2000) found a positive relationship between the use of BSC performance, whereas Ittner *et al.* (2003) found a negative relationship. In the public sector, performance management research has often been addressed as part of new public management (NPM) research (Hood, 1995; Gendron *et al.*, 2000). However, much of this literature has been concerned with critiques of NPM rather than empirical studies. Two recent surveys of performance management practice have been published: Lapsley and Wright (2004) in the UK and Carlin (2004) in Australia. However, there have been only a few empirical studies of performance management. These include Goddard (1992) who noted the importance of involving all stakeholders in the development of performance management models and the problems of power and conflict between them. Modell (2003, 2004) found resistance to imposed PMs and resulting conflict in the Norwegian health care sector and that implementation of goal-directed performance management models does not necessarily

reduce conflict and ambiguity. Collier (2006) and Guven-Uslu (2006) both noted the importance of relating PMs to strategic priorities in policing and health respectively. Robinson (2003) found that the way performance measurement had been used by government in Alberta had created accountability problems and that there was a discrepancy between what they intended and what they produced. Goddard (2005) noted the continuing reliance on accounting practices, particularly budgeting, rather than non financial PMs in UK local government. However, Modell (2004) has argued that public services may be moving from a financial control emphasis towards a multidimensional performance measurement model. Clearly further research is necessary concerning this possible transition.

Two studies have researched the BSC in public services. Aidemark (2001) looked at the BSC in a health care organization in Sweden. He found it was used more as a communication device between clinicians than a goal-directing device. McAdam and Walker (2003) investigated the use of the BSC within best value implementation in UK local government. They found that although the BSC played a key role, its implementation was problematic. In contrast to the case study research, Cavalluzzo and Ittner (2003) used a contingency theory approach to study the relationship between implementation factors, performance management systems and their outcomes in a survey of managers in US federal government. They found that implementation factors were important but that overall a complex relationship existed. Prior empirical, public sector research is mainly from outside the UK and is eclectic rather than systematic.

### Methodology

The research reported in this article looked at the relationship between strategic management, accounting and performance management systems. Organizations are complex, socially constructed phenomena and our research therefore used an interpretive approach. One interpretive methodology which has been found to work well by public sector accounting researchers is grounded theory (Parker and Roffey, 1997; Goddard, 2004, 2005). It is particularly helpful where theoretical understanding as well as empirical knowledge is sought. Various approaches to grounded theory have been identified (Locke, 2001) especially with respect to the extent to which prior theory is used and to which the research problem is defined. This study set broad

research questions to develop a nascent theory. The questions addressed such issues as how performance management systems are developed, perceived and used; what theoretical explanation can be developed to understand and explain the interrelationships between performance management systems, accounting practices and strategy; what differences in these phenomena exist between services and to what extent problems related to cause and effect, conflicting interests and ambiguity arise.

#### *Data Collection*

The study comprises four case studies: two in local government and two in higher education institutions. The two sectors were chosen as they represented the two ends of the spectrum of governance within the public sector. The local government cases were unitary authorities. In the latest Audit Commission comprehensive performance assessment (CPA) ratings one of the organizations was 'improving adequately' and demonstrating a 3-star overall performance and the other was 'improving well' and also 3 star.

The universities comprise one 'old' university and one 'new'. The old university has a long history of independence. The new university was established as an independent university in 1992 and was previously managed by the local authority. In general, old universities have placed more emphasis on research and new universities more on teaching. This was evident in the two selected cases' performance under the latest Research Assessment Exercise (RAE), which is a peer review of the research of all UK universities and rates each department on a six-point scale. The 'new' university obtained 24, 5 or 5\* ratings (the highest possible) out of 34 submitted departments, whereas the 'old' university obtained 1 out of 9 submissions. One of each type of university was chosen to explore the possibility of the different historical organizational arrangements having an effect on current performance management.

A series of interviews was undertaken for each case study. In the local government organizations, these included managers and members involved with the use of PMs, performance management, strategy and accounting systems. Data was collected from both central and service departments. A similar strategy of data collection from central and academic departments and schools was undertaken in the universities. In total some 80 interviews have been undertaken.

A description of PMs, performance

management, strategy and accounting practices in each organization was obtained, together with participants' perceptions of these phenomena and practices. Triangulation was undertaken by a document search of relevant publications and committee minutes, including performance management reports, financial reports, audit reports, and governance reports. Whenever possible, observation was also undertaken, including attendance at significant meetings.

#### *Data Analysis*

The data was analysed using the series of coding procedures suggested by Strauss and Corbin (1998), which comprises three different stages:

- Open coding, which produces a set of concepts which play a role in the case organization's life with respect to performance management, strategy and accounting.
- Axial coding, which concentrates on the relationships between performance management, strategy and accounting to produce a set of main categories. Reducing the number of categories at this stage allows for a higher level of abstraction to be reached.
- The final coding stage, selective coding, is the process of integrating and refining the grounded theory.

All categories are unified around a core category, which is the main theme of the research.

#### **Results**

Our results are presented here in two stages. A summary of the main issues are presented first. The principal components of the more abstract grounded theory follow, which explain the differences between organizations and establishes linkages between the sets of issues.

#### **Main Issues**

##### *Principal Differences*

The local authority case studies were characterized by highly centralized practices associated with strategic management, budgeting and performance management. These practices were, moreover, highly structured with all parts of the organization conforming to the corporate requirements. Perhaps the most striking difference between the local authorities and higher education was the extensive use of both strategic planning and of performance measures, with sophisticated hierarchical reporting procedures. In the two local government case

studies some 90 and 70 key PMs, respectively, were regularly reported to the executive boards. In addition, both organizations publish annually in excess of 300 PMs and internal departments often had yet more PMs of their own. In the higher education cases, around 15 PMs were regularly reported and few additional PMs used in the departments.

There were few differences between the local government case studies or intra-organizationally. In both cases, what might be termed a strong performance management culture had developed. All participants valued PMs and considered them an essential organizational tool. It was evident that external pressures, particularly from the Audit Commission's CPA and the best value initiatives, were very strong and influential. However, there were concerns over the imposition of specific measures and with the dysfunctional effects on strategy. This was caused by the perceived need to improve performance against imposed targets, even though the organization did not necessarily wish to pursue such a service strategy.

Higher education was characterized by relatively decentralized practices. There was little real strategic planning taking place. Such planning as there was was essentially annual and concerned with short-term budgeting. Relatively little use was made of performance measures, other than the RAE ratings and the national student survey. Strategic planning and performance management was evolutionary rather than prescribed. Little hierarchical reporting was found and there was a lack of formalized reporting loops. As one commentator put it, 'where in the past we have had a performance management issue...it goes further up the line, then I find it does tend to fizzle out somewhat'. This lack of formalized reporting systems and feedback loops, coupled with the collegiate culture of academia, had resulted in a performance management system which had no clear chain of authority or ownership. Other than the two broad measures referred to above, there was little or no influence from external PM setting. RAE scores did have significant influence on strategy in the case of the old university, but little effect in the new university where the national student survey appeared to be of greater importance. In both cases, only a weak performance culture existed. However, practices varied across the cases and even more within each organization across division and schools.

### *Commonalities*

There were some commonalities between the two sectors. In both, there was an increasing use of PMs and this was leading to increased centralization. PMs were also associated with a shift of power towards managers and away from councillors and academics. There was a sense that PMs were largely concerned with managers talking to managers (departments/division reporting to central managers, central managers reporting to external bodies). There was little evidence of PMs being used for reporting to other stakeholders in any meaningful way. There appeared to be a widening accountability gap emerging between managers and other stakeholders. This was particularly true in local government where the cabinet system was marginalizing back-bench councillors. Dysfunctions of PMs were still fairly prevalent, such as contradictory PMs, and there was evidence of managing to achieve a better PM rather than service outcomes. Gaming occurred in both sectors with examples found of manipulating PMs in local government to improve CPA scores and manipulation in higher education to improve RAE scores. Within higher education, academics outside the core executive groups seemed to have no knowledge or interest in the PMs and even within the executive group there was some disconnection.

### *Context*

There were three main contextual differences between the two settings:

- *Governance structures*: the existence of strong political accountability through elected councillors in local government had no equivalent in higher education. This has had a very significant affect on local government and created cultural differences between the two types of organization. Within higher education there was a view that they 'aren't a heavily line managed and therefore performance managed institution...it's not suitable for this type of organization'. The organizational culture was perceived as being, 'kind of unique...a collegiate process and a school of academics'. This had led to a performance management system that was not embedded in the organizational culture.
- *The importance of external mechanisms*: CPA in local government and the RAE/national student survey in higher education. In both cases these external mechanisms have had an impact on the cultures and practices of the organizations, but this was much stronger in local government.

• *The range of services provided*: the broad range of significantly different services provided by local government required an increased number of PMs compared to higher education. However, the differences in market influence between the two sectors did not emerge as important in relation to PMs. Gaming occurred in both sector as discussed above.

#### *Nature of Measures*

Issues about the measures were more prevalent in local government, probably because PMs were more heavily used in this area. These issues are also well established in the academic literature. The first concerns the emphasis placed on quantifiable, measurable, process-related PMs associated with a paucity of outcome indicators. Within higher education, strategies had been developed which were, 'without definable measurement'. They were 'a bit woolly' and therefore it was difficult 'to put your finger on exactly how they measure[d] performance'. Using the BSC terminology, in both sectors there was predominance of internal business process measures. Indeed, a crude analysis of the PMs revealed that around 75% of PMs were of this type. Consequently, there were relatively few financial or customer-related measures and almost no innovation and growth measures. This resulted in a very narrow, internally-biased focus of performance management in all the organizations studies. Similarly, there were relatively few 'local' indicators in either local government or higher education. There were concerns about whether achieving PM targets achieved better services in both areas. The lack of clear measurement guidelines for internally developed strategies within higher education resulted in a view that, 'nobody knows how its measured and there's no sanctions if its not measured'.

#### *Management Issues*

A set of issues concerning the management of PMs also emerged in all case studies. The first concerned a lack of connectivity between PMs and budgets/financial consequences. There tended to be a separate reporting process for PMs and for financial control. In both areas PMs were mainly being used for control purposes rather than resource and policy planning. There was little evidence of the interrelationship between PMs being used in the strategic planning of the organizations as exhorted by the 'aeroplane cockpit' approach of the BSC. The excessive number of PMs in local government was a concern of several

interviewees and problems of bounded rationality were evident.

#### **Emerging Grounded Theory**

Despite local government and higher education both being in the public sector, they have adopted vastly different approaches to performance management. In order to understand the process by which PMs were being managed in all the case studies and to explain these differences, we developed a grounded theory from the data.

The most important phenomenon which emerged from the study was way the organizations' participants have come to view how they will be managed. This has been labelled the managerial 'worldview' for ease of reference. It emerges from external influences, past experiences and practices, and participants' own worldviews. The managerial worldview sets the agenda for management in the organization in terms of what the main purposes and priorities of management are (these might be to control the organization, to ensure professional competence and quality, the attitude to accountability, the importance of PMs etc.). It also determines how this purpose is to be achieved. In the case studies this ranged from a highly centralized, structured and formalized approach in local government (a bureaucratic approach) to the decentralized and informal approach in higher education (an amorphous approach).

The managerial worldview is in constant evolutionary change, subject to changes in contextual factors, participants and practices. It is also subject to power struggles within the organization between stakeholders (politicians and managers, central managers and departmental managers, managers and professionals). The outcome of these struggles is important in determining which views come to dominate or most influence the worldview.

The managerial worldview has a major influence on how and what management practices will be undertaken, and permeates the way in which these practices are perceived and implemented. In this study, two phenomena emerged to understand these practices: 'strategizing' and 'performancing'. Strategizing refers to the mindsets of participants with respect to both strategy and strategic management and to the associated practices. It encapsulates perceptions about the meaning of strategy within the organizational and disciplinary context. Depending on the sectoral, departmental and disciplinary setting, strategizing takes different

forms. For instance in local government, strategy was perceived to be extremely important to managers who generally adopted a long-term focus and objectives, whereas in higher education (and local government members) strategy was less important and they adopted a relatively short-term view of strategic management with annual planning and budgeting being the main strategic planning concern. Moreover, strategizing practices conformed to the managerial worldview of the organization. In local government cases they were highly bureaucratic in both design and operation, and in higher education they were amorphous.

Performancing refers to the mindsets of participants about performance management and performance measures. The variations mirrored those found with strategizing, with performance management being perceived as extremely important and useful by local government managers but far less so by higher education managers and local government members. Similarly performancing practices conformed to the managerial worldview of the organizations being highly bureaucratic in local government and amorphous in higher education as outlined above. It was also evident, particularly in local government, that performance management practices were also reinforcing and/or changing the managerial worldview. The performance management practices had developed significantly over a period of several years and managers and members had changed their views from one of caution and even cynicism to one of positive support.

The managerial worldview and associated strategizing and performancing is highly contextual and subject to external influences. In local government, the strongest influence was central government. Central government requirements, like CPA, have resulted in a strong emphasis on accountability in general and political accountability specifically. The accountability culture is conducive to performance management as it is familiar with reporting accounts between different elements of the organization. Political accountability is conducive to PMs because publicly reported CPA scores may well influence voting behaviour, or, at least, may be perceived to do so. Consequently, there was a strongly perceived need to control CPA by use of PMs and centralized practices were adopted that ensured sophisticated and structured collection and reporting of data throughout the organizations.

In higher education, external influences

were far less apparent and although central government has imposed some important measures (such as RAE) none require frequent, regular reporting practices as in local government. In higher education the managerial worldview was more a product of organizational history where the academic and disciplinary, as opposed to managerial, culture was still very strong (although variable within each organization). Values such as academic freedom, independence of discipline and suspicion of management held sway. The managerial worldview was not therefore conducive to sophisticated strategizing and performancing. Consequently attempts to implement such systems have failed. The perceived need for performance management in higher education was evident mostly in the central departments and seemed to emanate from a broader 'ideological' agenda which encouraged the use of PMs. However, the response to performance in each division varied. For instance the consequences of non-performance varied from being ignored to job loss. Perceived academic reality was a strong influence and academic freedom did prove an obstacle to performance management in some divisions, resulting in differences between academic and service divisions. Some differences were also observed between the two universities studied, with the new university being marginally more accepting of performance management reporting. This may be due to its historical experience within local government management.

Strategizing and performancing were closely related in terms of being conducive to the broader managerial worldview and organizations' strategies were linked to the PMs. However, the notion of using PMs to inform strategic decision-making and resource allocation, prevalent in the BSC literature, did not emerge strongly. The exceptions were that strategic decisions were made to improve CPA and RAE scores. However, in general, the PMs were used and perceived as control mechanisms. Accounting was implicated in strategizing and performancing principally through the budgeting process. However the perceptions and practice of budgeting depended on the divisional strategizing. The budget was either seen as the main long-term, strategic planning mechanism or as a financial planning tool merely for the coming year. The links between the budgeting process and PMs were not close, with separate reporting mechanisms and processes existing for each. The interrelationship between financial and other

perspectives, again emphasised in the BSC literature, was not strong.

Despite the significant differences in performance management practices between the two sectors, participants from both said they were dissatisfied with the practices currently used. In local government, this dissatisfaction was primarily with the imposition of the measures by central government and with the dysfunctional aspects of some of these measures. There was, however, a general satisfaction with use of PMs, particularly if they are locally determined. In higher education, the dissatisfaction was more fundamental as PMs were generally seen as 'managerialist' and in conflict with the academic worldview.

### Lessons for Practice and Suggestions for Improvement

Perhaps the most significant lesson for policy-makers is the vital importance of ensuring that the managerial worldview in the organization is conducive to performance management. This takes time and two principal approaches emerged to ensure its achievement:

- The first is the development of an organizational culture. This can be achieved by senior managers emphasising the importance of performance management through agenda management and verbal and written communication. It is important that all organizational participants are aware that performance management is valued and prized.
- The second approach is to establish performance practices such as the regular reporting of as reasonable number of measures, as such practices eventually become habitual and supportive of the culture.

In local government, intensive external imposition of PMs has changed practice and assisted in developing a performance management culture. The strong sense of accountability, associated with tradition of political accountability, has strengthened this worldview. Locally-determined PMs can now emerge naturally and will be far more effective in improving services than centrally imposed PMs. Indeed, the dysfunctions of externally-imposed PMs are already producing cynicism and if not addressed risk destroying the performance culture which has been developed. A key question for the regulatory bodies is whether they want to control local government services by detailed control of PMs, or whether

they wish to build on the success of changing practices and rely on ensuring that strong and effective performance management exists.

In higher education, the development of a performance management valuing worldview is unlikely to be achieved quickly by imposition as the existing culture values independent decentralization and there is little sense of accountability. The use of metric-based RAE, more akin to the CPA system in local government, will be more effective in establishing such a worldview than the existing RAE peer assessments. This is because universities will be more likely to impose internal practices centred on PMs and it is evident from local government that changing practices do affect worldview. However, care will have to be taken to minimize the inevitable dysfunctions. The development of a performance management worldview is crucial to effective performance management. Such culture change will be best achieved by encouraging spread of good practice, using such approaches as 'beacon' schemes and audit reports on performance management, rather than imposing a raft of PMs.

More generally, public sector organizations need to be aware of the vital importance of context to the development of a worldview conducive to performance management. Changing organizational practices have a role to play in this development but only if there is an awareness of the importance of dysfunctions. As with local government, regulatory bodies should seek to ensure effective performance management rather than control by PMs. There are also a number of issues concerning the nature of the PMs and of their management, outlined above, which need to be taken more seriously.

This article has highlighted the importance of organizational management of performance measures. Such management is varied and complex but has a crucial affect on the use of performance measures. Culture is an extremely important phenomenon in this respect and it is itself complex and heterogenous. Only by understanding how and why performance management works will it be possible to improve the delivery of our public services. ■

### References

- Aidemark, L.-G. (2001), The meaning of balanced scorecards in the health care organization. *Financial Accountability and Management*, 17, 1, pp. 23–40.
- Carlin, T. (2004), Output-based management and the management of performance.

- Management Accounting Research*, 15, pp. 267–283.
- Cavalluzzo, K. S. and Ittner, C. D. (2003), Implementing performance measurement innovations: evidence from government. *Accounting, Organizations and Society*, 29, pp. 243–267.
- Collier, P. M. (2006), In search of purpose and priorities: police performance indicators in England and Wales. *Public Money & Management*, 26, 3, pp. 165–172.
- Gendron, Y., Cooper, D. J. and Townley, B. (2001), In the name of accountability—state auditing, independence and new public management. *Accounting, Auditing & Accountability Journal*, 14, 3, pp. 278–310.
- Goddard, A. R. (1992), Perspectives on management control in a multiple agency, community service. *Financial Accountability and Management*, 8, 2, pp. 115–128.
- Goddard, A. R. and Powell, J. R. (1994), Accountability and accounting: using naturalistic methodology to enhance organizational control. *Accounting, Auditing, and Accountability Journal*, 7, 2, pp. 50–69.
- Goddard A. R. (2005), Accounting and NPM in UK local government—contributions towards accountability. *Financial Accountability and Management*, 21, 2, pp. 191–218.
- Güven-Uslu, P. (2006), Uses of performance metrics in clinical and managerial networks. *Public Money & Management*, 26, 2, pp. 95–100.
- Hood, C. (1995), The new public management in the 1980s: variations on a theme. *Accounting, Organizations and Society*, 20, 2/3, pp. 93–110.
- Hoque, Z. and James, W. (2000), Linking balanced scorecard measures to size and market factors: impact on organizational performance. *Journal of Management Accounting Research*, 12, pp. 1–17.
- Ittner, C., Larcker, D. F. and Randall, T. (2003), Performance implications of strategic performance measures in financial services firms. *Accounting, Organizations and Society*, 28, pp. 715–741.
- Kaplan, R. S. and Norton, D. P. (1992), The balanced scorecard—measures that drive performance. *Harvard Business Review* (January–February), pp. 71–79.
- Lapsley, I. and Wright, E. (2004), The diffusion of management accounting innovations in the public sector: a research agenda. *Management Accounting Research*, 15, pp. 355–374.
- Locke, K. (2001), *Grounded Theory in Management Research* (Sage Publications, London).
- McAdam, R. and Walker, T. (2003), An inquiry into balanced scorecards within best value implementation in UK local government. *Public Administration*, 8, 4, pp. 873–892.
- Modell, S. (2003), Goals versus institutions: the development of performance measurement in the Swedish university sector. *Management Accounting Research*, 14, pp. 333–359.
- Modell, S. (2004), Performance measurement myths in the public sector: a research note. *Financial Accountability and Management*, 20, 1, pp. 39–55.
- Parker, L. D. and Roffey, B. H. (1997), Methodological themes: back to the drawing board. *Accounting, Auditing and Accountability*, 10, 2, pp. 212–247.
- Robinson, P. (2003), Government accountability and performance measurement. *Critical Perspectives on Accounting*, 14, pp. 171–186.
- Strauss, A. and Corbin, J. (1998), *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*, 2nd edn (Sage, Thousand Oaks).